Universidade de Aveiro Departamento de Química
2014

Cláudia Manuela Mesquita da Rocha

Assinatura metabólica do cancro do pulmão: estudo metabolómico de tecidos e biofluidos humanos

Metabolic signature of lung cancer: a metabolomic study of human tissues and biofluids

Universidade de Aveiro Departamento de Química
2014

Cláudia Manuela Mesquita da Rocha

Assinatura metabólica do cancro do pulmão: estudo metabolómico de tecidos e biofluidos humanos

Metabolic signature of lung cancer: a metabolomic study of human tissues and biofluids

Tese apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Química, realizada sob a orientação científica da Doutora Iola Melissa Fernandes Duarte, Investigadora Auxiliar do Laboratório Associado Centro de Investigação em Materiais Cerâmicos e Compósitos (CICECO), Departamento de Química da Universidade de Aveiro.

Apoio financeiro da Fundação para a Ciência e Tecnologia (FCT) - bolsa de investigação FCT SFRH/BD/63430/2009 financiada pelo Programa Operacional Potencial Humano (POPH) e projetos FCT/PTDC/QUI/68017/2006 e FCOMP-01-0124-FEDER-037271 (Ref. FCT PEst-C/CTM/LA0011/2013) financiados pelo Fundo Europeu de Desenvolvimento Regional (FEDER) através do Programa Operacional Fatores de Competitividade (COMPETE); da Universidade de Aveiro - Centro de Investigação em Materiais Cerâmicos e Compósitos (CICECO); do Centro de Investigação em Meio-Ambiente, Genética e Oncobiologia (CIMAGO), da Faculdade de Medicina da Universidade de Coimbra; da Liga Portuguesa Contra o Cancro. Agradecemos ainda à Rede Nacional de RMN (RNRMN), suportada com fundos da FCT, à empresa Bruker BioSpin GmbH e ao Imperial College London (Department of Surgery and Cancer), Londres, Inglaterra.

Dedico aos meus pais, João e João Pedro

o júri

presidente

Prof. Doutor Paulo Jorge de Melo Matias Faria de Vila Real
Professor Catedrático do Departamento de Engenharia Civil da Universidade de Aveiro

Prof. Doutora Maria Helena Dias Santos
Professora Catedrática Aposentada do Instituto de Tecnologia Química e Biológica da Universidade Nova de Lisboa

Doutor Tiago Brandão Rodrigues
Investigador Auxiliar da Universidade de Cambridge – Cancer Reasearch UK Cambridge Institute, Reino Unido

Prof. Doutor Carlos Manuel Silva Robalo Cordeiro
Professor Associado com Agregação da Faculdade de Medicina da Universidade de Coimbra, Presidente da Sociedade Portuguesa de Pneumologia

Prof. Doutora Ana Maria Pissarra Coelho Gil
Professora Associada com Agregação do Departamento de Química da Universidade de Aveiro

Doutora Iola Melissa Fernandes Duarte
Investigadora Auxiliar do Laboratório Associado CICECO, Departamento de Química da Universidade de Aveiro

agradecimentos

Em primeiro lugar, agradeço a toda a equipa envolvida neste projeto, sem a qual, tal como foi constituída, não teria sido possível o sucesso do mesmo.

Assim, agradeço à pessoa que me orientou e ajudou durante mais de seis anos, a Doutora Iola Duarte. Agradeço-lhe pela oportunidade de ter participado num projeto desafiante em todos os momentos. Agradeço ainda os momentos de aprendizagem científica, ética profissional e pessoal.

Gostaria de agradecer à equipa do Instituto de Anatomia Patológica da Faculdade de Medicina da Universidade de Coimbra, em especial à Prof. Doutora Lina Carvalho, ao Dr. Vítor Sousa e à Drª Ana Gomes pelo incansável empenho na coordenação da recolha e processamento das amostras dos doentes e por todo o apoio prestado; à equipa do Centro de Cirurgia Cardiotorácica do Centro Hospitalar e Universitário de Coimbra, em particular ao Dr. João Bernardo; e, também, à Prof. Doutora Isabel M. Carreira e à Drª Joana Melo do Laboratório de Citogenética e Genómica da Faculdade de Medicina da Universidade de Coimbra pelos contínuos incentivos.

Agradeço à empresa INDASA S.A., na pessoa do Engº Nelson Ramos, pela colaboração e total disponibilidade, facilitando o acesso à recolha de amostras de voluntários saudáveis.

O meu profundo agradecimento a todos os voluntários, doentes e saudáveis, que participaram neste estudo, esperando com isso ter contribuído para algum avanço no conhecimento da doença. Sem eles não teria sido possível.

Agradeço ao grupo de metabolómica da Universidade de Aveiro: à Prof. Doutora Ana Gil, ao Doutor António Barros e ao Doutor Brian Goodfelllow, com os quais muito aprendi nas diversificadas discussões em contexto de trabalho, mas não só.

Agradeço às pessoas que no Imperial College of London me permitiram ter contacto com novas técnicas, num ambiente muito estimulante, e me ajudaram no trabalho de espetrometria de massa: a Doutora Maria Gomez-Romero, o Doutor Mathew Lewis e a Prof. Doutora Elaine Holmes. Agradeço ainda à Bruker Biospin, na pessoa do Doutor Manfred Spraul, em particular pelo acesso a software especializado e à base de dados de RMN.

Gostaria de agradecer a todo os meus amigos, aos presentes e aos ausentes. Em especial, aos amigos que fiz durante este trajeto: ao Gonçalo Graça, à Inês Lamego, à Joana Carrola, à Joana Marques, à Joana Pinto, ao João Rodrigues, ao Ricardo Mendes, ao Sérgio Vilela e à Susana Aveiro, sempre incansáveis na boa disposição, no companheirismo e na amizade. Tornaram sempre tudo mais fácil e melhor, com certeza.

Por fim gostaria de agradecer profundamente a toda a minha família: pais, irmãos, cunhados, sobrinhos, sogros, avós, tios e primos; em particular, claro, agradeço ao João e ao João Pedro, em quem tantas vezes procurei inspiração, e encontrei.

palavras-chave

cancro do pulmão, metabolómica, perfil metabólico, metabolismo do cancro, tecido tumoral, plasma sanguíneo, urina, espetroscopia de ressonância magnética nuclear (RMN), RMN de alta resolução com rotação no ângulo mágico, cromatografia líquida de ultra-eficiência acoplada a espetrometria de massa (UPLC-MS), análise multivariada

resumo

A presente tese reporta a aplicação da metabolómica ao estudo de tecidos e biofluidos humanos (plasma sanguíneo e urina), com o intuito de caracterizar a assinatura metabólica do cancro pulmonar primário. No Capítulo 1, apresenta-se uma breve introdução sobre a epidemiologia e a patogénese deste tipo de cancro, bem como um sumário das principais alterações metabólicas tipicamente associadas ao cancro em geral. Descreve-se ainda a abordagem metabolómica, nomeadamente os métodos analíticos e estatísticos utilizados, assim como o estado da arte da sua aplicação em estudos clínicos do cancro do pulmão.

No Capítulo 2, apresentam-se os detalhes experimentais deste trabalho, no que diz respeito ao grupo de indivíduos envolvidos, à colheita e análise das amostras e ao posterior tratamento dos dados.

O Capítulo 3 descreve a caracterização metabólica de tecidos do pulmão (de 56 doentes) por espetroscopia de Ressonância Magnética Nuclear (RMN) de alta resolução com rotação no ângulo mágico. Após a otimização cuidada das condições de aquisição e a identificação detalhada dos sinais espetrais (mais de 50 metabolitos identificados), os perfis metabólicos dos tumores e dos tecidos adjacentes não envolvidos (controlos) foram comparados por análise multivariada, tendo sido discriminados com uma exatidão de 97%. Os metabolitos que mais significativamente contribuíram para esta diferenciação foram: glucose e acetato (diminuídos nos tumores), lactato, alanina, glutamato, GSH, taurina, creatina, fosfocolina, glicerofosfocolina, fosfoetanolamina, nucleótidos de uracilo e péptidos (aumentados nos tumores). Algumas destas variações corroboraram alterações típicas do metabolismo do cancro (e.g., glicólise e glutaminólise aumentadas), enquanto outras sugeriram novas pistas sobre a possível relevância de processos como a proteção antioxidante e a degradação proteica. Um outro resultado novo e importante descrito neste capítulo foi a dependência da assinatura metabólica em relação ao tipo histológico do tumor. Enquanto as principais alterações observadas nos adenocarcinomas (AdC) se relacionaram com o metabolismo fosfolipídico e proteico, os carcinomas de células escamosas (SqCC) apresentaram perfis glicolíticos e glutaminolíticos mais pronunciados, sendo possível construir um modelo válido para a discriminação destes subtipos.

No Capítulo 4, apresenta-se o estudo metabolómico por RMN de plasma sanguíneo de mais de 100 doentes e quase 100 controlos saudáveis, do qual resultou um modelo multivariado com uma taxa de classificação de 87%. A distinção entre os grupos foi feita essencialmente com base nos níveis de lactato, piruvato, acetoacetato, lipoproteínas LDL+VLDL e glicoproteínas (aumentados nos doentes), juntamente com os níveis de glutamina, histidina, valina, metanol, lipoproteínas HDL e dois compostos não identificados (diminuídos nos doentes). Estas variações foram detetadas desde os estádios iniciais da doença e a magnitude de algumas delas dependeu do tipo histológico, embora não permitindo discriminar AdC de SqCC. Para além disso, mostra-se neste capítulo que o desequilíbrio dos grupos controlo e cancro em termos da idade dos indivíduos poderá ter alguma influência nos resultados, e apresenta-se uma tentativa exploratória de validação externa, que resultou numa taxa de classificação de 85%.

O estudo por RMN do perfil metabólico da urina dos doentes com cancro do pulmão e dos controlos é apresentado no Capítulo 5. Comparativamente ao plasma, o modelo construído com os perfis urinários apresentou uma taxa de

classificação superior (97%). Após uma avaliação cuidada da possível influência do género, idade e hábitos tabágicos, um conjunto de 19 metabolitos foi proposto como estando relacionado com a doença (incluindo 3 compostos desconhecidos e 6 parcialmente identificados como metabolitos *N*-acetilados). Tal como no caso do plasma, estas variações foram detetadas em doentes no estádio inicial e mostraram alguma dependência em relação ao tipo histológico, obtendo-se um modelo válido para a discriminação AdC vs. SqCC, ainda que com um poder preditivo modesto. Para além disso, o teste preliminar de validação externa revelou 100% de sensibilidade e 90% de especificidade, o que é um resultado bastante promissor em termos da potencial utilização dos perfis urinários em aplicações clínicas futuras.

No Capitulo 6, descreve-se a caracterização dos perfis metabólicos da urina (de um subgrupo de indivíduos) por cromatografia líquida de ultra-eficiência acoplada a espetrometria de massa (UPLC-MS). Embora não avançando muito na identificação estrutural de possíveis marcadores, este estudo reforçou o valor diagnóstico da urina, já que os modelos multivariados resultantes apresentaram taxa de classificação e poder preditivo elevados.

Finalmente, no Capítulo 7, apresentam-se as principais conclusões deste trabalho, realçando o contributo da metabolómica integrada de tecidos e biofluidos para a compreensão do metabolismo alterado do cancro do pulmão e para a deteção de novos perfis marcadores com valor diagnóstico.

keywords

lung cancer, metabolomics, metabolic profile, cancer metabolism, tumour tissue, blood plasma, urine, nuclear magnetic resonance (NMR) spectroscopy, high resolution magic angle spinning (HRMAS), ultra-performance liquid chromatography-mass spectrometry (UPLC-MS), multivariate analysis (MVA)

abstract

This thesis reports the application of metabolomics to human tissues and biofluids (blood plasma and urine) to unveil the metabolic signature of primary lung cancer. In Chapter 1, a brief introduction on lung cancer epidemiology and pathogenesis, together with a review of the main metabolic dysregulations known to be associated with cancer, is presented. The metabolomics approach is also described, addressing the analytical and statistical methods employed, as well as the current state of the art on its application to clinical lung cancer studies.

Chapter 2 provides the experimental details of this work, in regard to the subjects enrolled, sample collection and analysis, and data processing.

In Chapter 3, the metabolic characterization of intact lung tissues (from 56 patients) by proton High Resolution Magic Angle Spinning (HRMAS) Nuclear Magnetic Resonance (NMR) spectroscopy is described. After careful assessment of acquisition conditions and thorough spectral assignment (over 50 metabolites identified), the metabolic profiles of tumour and adjacent control tissues were compared through multivariate analysis. The two tissue classes could be discriminated with 97% accuracy, with 13 metabolites significantly accounting for this discrimination: glucose and acetate (depleted in tumours), together with lactate, alanine, glutamate, GSH, taurine, creatine, phosphocholine, glycerophosphocholine, phosphoethanolamine, uracil nucleotides and peptides (increased in tumours). Some of these variations corroborated typical features of cancer metabolism (e.g., upregulated glycolysis and glutaminolysis), while others suggested less known pathways (e.g., antioxidant protection, protein degradation) to play important roles. Another major and novel finding described in this chapter was the dependence of this metabolic signature on tumour histological subtype. While main alterations in adenocarcinomas (AdC) related to phospholipid and protein metabolisms, squamous cell carcinomas (SqCC) were found to have stronger glycolytic and glutaminolytic profiles, making it possible to build a valid classification model to discriminate these two subtypes.

Chapter 4 reports the NMR metabolomic study of blood plasma from over 100 patients and near 100 healthy controls, the multivariate model built having afforded a classification rate of 87%. The two groups were found to differ significantly in the levels of lactate, pyruvate, acetoacetate, LDL+VLDL lipoproteins and glycoproteins (increased in patients), together with glutamine, histidine, valine, methanol, HDL lipoproteins and two unassigned compounds (decreased in patients). Interestingly, these variations were detected from initial disease stages and the magnitude of some of them depended on the histological type, although not allowing AdC vs. SqCC discrimination. Moreover, it is shown in this chapter that age mismatch between control and cancer groups could not be ruled out as a possible confounding factor, and exploratory external validation afforded a classification rate of 85%.

The NMR profiling of urine from lung cancer patients and healthy controls is presented in Chapter 5. Compared to plasma, the classification model built with urinary profiles resulted in a superior classification rate (97%). After careful assessment of possible bias from gender, age and smoking habits, a set of 19 metabolites was proposed to be cancer-related (out of which 3 were unknowns and 6 were partially identified as *N*-acetylated metabolites). As for plasma, these variations were detected regardless of disease stage and showed some dependency on histological subtype, the AdC vs. SqCC model built showing modest predictive power. In addition, preliminary external validation of the urine-based classification model afforded 100% sensitivity and 90% specificity, which are exciting results in terms of potential for future clinical application.

Chapter 6 describes the analysis of urine from a subset of patients by a different profiling technique, namely, Ultra-Performance Liquid Chromatography coupled to Mass Spectrometry (UPLC-MS). Although the identification of discriminant metabolites was very limited, multivariate models showed high classification rate and predictive power, thus reinforcing the value of urine in the context of lung cancer diagnosis. Finally, the main conclusions of this thesis are presented in Chapter 7, highlighting the potential of integrated metabolomics of tissues and biofluids to improve current understanding of lung cancer altered metabolism and to reveal new marker profiles with diagnostic value.

List of publications derived from the work presented in this thesis

Rocha C.M., Barros A.S., Gil A.M., Goodfellow B.J., Bernardo J., Carvalho L., Sousa V., Carreira I.M., Melo J.B., Humper E., Spraul M., Duarte I.F., 2010. Metabolic profiling of human lung cancer tissue by high resolution magic angle spinning (HRMAS) ^{1}H NMR spectroscopy. *Journal of Proteome Research*, 9(1), pp.319–32.

Duarte I.F., Rocha C.M., Barros A.S., Gil A.M., Goodfellow B.J., Carreira I.M., Bernardo J., Gomes A., Sousa V., Carvalho L., 2010. Can Nuclear Magnetic Resonance (NMR) spectroscopy reveal different metabolic signatures for lung tumours? *Virchows Archiv*, 457(6), pp.715-25.

Carrola J., Rocha C.M., Barros A.S., Gil A.M., Goodfellow B.J., Carreira I.M., Bernardo J., Gomes A., Sousa V., Carvalho L., Duarte I.F., 2011. Metabolic signatures of lung cancer in biofluids: NMR-based metabonomics of urine. *Journal of Proteome Research*, 10(1), pp.221-30.

Rocha C.M., Carrola J., Barros A.S., Gil A.M., Goodfellow B.J., Carreira I.M., Bernardo J., Gomes A., Sousa V., Carvalho L., Duarte I.F., 2011. Metabolic signatures of lung cancer in biofluids: NMR-based metabonomics of blood plasma. *Journal of Proteome Research*, 10(9), pp.4314-24.

Rocha C.M., Barros A.S., Goodfellow B.J., Carreira I.M., Gomes A., Sousa V., Bernardo J., Carvalho L., Gil A.M., Duarte I.F., 2014. NMR metabolomics of human lung tumours reveals distinct metabolic signatures for adenocarcinoma and squamous cell carcinoma. *Carcinogenesis*, in press.

Manuscripts in preparation

Rocha C.M., Gomez-Romero M., Lewis M.R., Barros A.S., Goodfellow B.J., Carreira I.M., Gomes A., Bernardo J., Carvalho L., Gil A.M., Holmes, E., Duarte I.F. Potential of urine profiling by UPLC-MS in the diagnosis of lung cancer.

Rocha C.M., Barros A.S., Goodfellow B.J., Carreira I.M., Gomes A., Sousa V., Bernardo J., Carvalho L., Gil A.M., Duarte I.F. Integrative tissue and biofluid metabolomics reveals metabolic dysregulations in lung cancer with potential diagnostic value.

CONTENTS

ABBREVIATIONS AND ACRONYMS

1D	one-dimensional
2D	two-dimensional
2DG	2-deoxyglucose
3PG	3-phosphoglycerate
3PO	3-(3-pyridinyl)-1-(4-pyridinyl)-2-propen-1-one
5FU	5-fluorouracil
ACC	acetyl-CoA carboxylase
ACCP	American College of Chest Physicians
ACL	ATP-citrate lyase
AdC	adenocarcinoma
ADP	adenosine diphosphate
AKT	protein kinase B
ALT	alanine aminotransferase
AMP	adenosine monophosphate
APCI	atmospheric pressure chemical ionisation
AQC	acquisition time
ASqCC	adenosquamous cell carcinoma
ATP	adenosine triphosphate
BAC	bronchoalveolar carcinoma
BALF	bronchoalveolar lavage fluid
BC	breast cancer
BMI	body mass index
CA	canonical analysis
CAF	cancer-associated fibroblast
CC	colon cancer
CDP	cytidine diphosphate
CE	capillary electrophoresis
CAE	carcinoembryonic antigen
ChoK	choline kinase
CI	chemical ionisation
CK	creatine kinase
CoA	coenzyme A
COPD	chronic obstructive pulmonary disease
COSY	correlation spectroscopy
COX	cyclo oxygenase
CPMG	Carr-Purcell-Meiboom-Gill
CR	classification rate
CRC	colorectal cancer
CSA	chemical shift anisotropy
CT	computed tomography
CV	coefficient of variation
CYFRA	serum cytokeratin fragment

Cys-LT	cysteinyl leukotriene
DESI	desorption electrospray ionisation
DFA	discriminant factor analysis
DGDP	deoxyguanosine diphosphate
DHAP	dehydroxyacetone phosphate
DIMS	direct infusion mass spectrometry
DNA	deoxyribonucleic acid
DNP	dynamic nuclear polarization
EBC	exhaled breath condensate
EGFR	epidermal growth factor receptor
EI	electron ionisation
EIA	enzyme immunoassay
EIC	extracted ion chromatogram
EMA	epithelial membrane antigen
ESI	electrospray ionisation
ESMO	European Society for Medical Oncology
F1,6BP	fructose 1,6-biphosphate
F6P	fructose 6-phosphate
FA	fatty acid
FAS	fatty acid synthase
FB	flexible bronchoscopy
FDG	^{18}fluorodeoxyglucose
FID	free induction decay
FN	false negative
FNAB	fine needle aspiration bronchoscopy
FP	false positive
FPR	false positive rate
FT	Fourier transform
FTICR	Fourier transform ion cyclotron resonance
G6PD	glucose-6-phosphate dehydrogenase
GA3P	glyceraldehydes-3-phosphate
GABA	gamma-aminobutyric acid
GAN	group aggregating normalization
GC	gas chromatography
GCA	gastric cancer
GLS	glutaminase
GLUT	glucose transporter
GMP	guanosine monophosphate
GPC	glycerophosphocholine
GPE	glycerophosphoethanolamine
GRP	gastrin-releasing peptide
GSH	reduced glutathione
GTP	guanosine triphosphate
H&E	hematoxylin and eosin

HCA	hierarchical analysis
HDL	high-density lipoprotein
HETCOR	heteronuclear correlation
HET-STOCSY	heteronuclear statistical total correlation spectroscopy
HIF	hypoxia inducible factor
HILIC	hydrophilic interaction chromatography
HK	hexokinase
HNC	head and neck cancer
HPLC	high performance liquid chromatography
HRMAS	high-resolution magic angle spinning
HSA	human serum albumin
HSQC	heteronuclear single-quantum coherence
HSS	high strength silica
HS-SPME	headspace solid-phase microextraction
INEPT	insensitive nuclei enhancement by polarization transfer
JRES	J-resolved
KRAS	Kirsten rat sarcoma
LASSO	least absolute shrinkage and selection operator
LB	line broadening
LC	liquid chromatography
LCC	large cell carcinoma
LDCT	low-dose computed tomography
LDH	lactate dehydrogenase
LDL	low-density lipoprotein
L-DON	6-diazo-5-oxo-L-norleucine
LED	longitudinal encoded-decoded
LOESS	locally weighted scatter plot smoothing
LPC	lysophosphatidylcholine
LTB4	leukotriene B4
LV	latent variable
MAPK	mitogen-activated protein kinase
MAS	magic angle spinning
MCCV	Monte Carlo cross validation
ME	malic enzyme
MEK	threonine and tyrosine recognition kinase
MLEV	Malcolm Levitt sequence
MLR	multiple logistic regression
MRI	magnetic resonance imaging
mRNA	microribonucleic acid
MRS	magnetic resonance spectroscopy
MRSI	magnetic resonance spectroscopy imaging
MS	mass spectrometry
mTOR	serine/threonine protein kinase
MVA	multivariate analysis

Mw	molecular weight
NAA	*N*-acetyl aspartate
NAD	nicotinamide adenine dinucleotide
NADPH	nicotinamide adenine dinucleotide phosphate
NCAM	neural cell adhesion molecule
NLST	National Lung Screening Trial
Neu5Ac	*N*-acetylneuraminic acid
NMR	nuclear magnetic resonance
NOESY	nuclear Overhouser effect spectroscopy
NS	number of scans
NSCLC	non-small cell lung cancer
NSE	neuron-specific enolase
OPLS	orthogonal projection to latent structures
OPLS-DA	orthogonal projection to latent structures – discriminant analysis
PC	phosphocholine
PCa	prostate cancer
PCA	principal component analysis
PCs	principal components
PDK	pyruvate dehydrogenase kinase
PE	phosphoethanolamine
PEP	phosphoenolpyruvate
PET	positron emission tomography
PFAA	plasma free amino acids
PFK	phosphofructokinase
PI3K	phosphatidylinositol 3-kinase
PK	pyruvate kinase
PLS	partial least squares
PLS-DA	partial least squares – discriminant analysis
PPP	pentose phosphate pathway
PQN	probabilistic quotient normalisation
PS	phosphatidylserine
PtdCho	phosphatidylcholine
PTE	phosphatidylethanolamine
QC	quality control
R5P	ribose-5-phosphate
RAS	rat sarcoma
RD	relaxation delay
RF	radiofrequency
RMN	ressonância magnética nuclear
RNA	ribonucleic acid
ROC	receiver operating characteristic
ROS	reactive oxygen species
RP	reverse phase
RSPA	recursive segment-wise peak

RT	retention time
SAM	*S*-adenosyl methionine
SC	sarcomatoid carcinoma
SCLC	small-cell lung cancer
SHY	statistical heterospectroscopy
SIRM	stable isotope resolved metabolomics
SM	sphingomyelin
SqCC	squamous cell carcinoma
SR	spinning rate
SREBP	sterol regulatory element binding protein
SSB	spinning side band
STOCSY	statistical total correlation spectroscopy
SW	spectral width
TCA	tricarboxylic acid
TD	time domain
THF	tetrahydropholate
TIC	total ion chromatogram
TK	tyrosine kinase
TKI	tyrosine kinase inhibitor
TKT	transketolase
TMAO	trimethylamine-*N*-oxide
TMS	tetramethylsilane
TN	true negative
TNA	transthoracic needle aspiration
TNM	tumour node metastasis
TOCSY	total correlation spectroscopy
ToF	time of flight
TP	true positive
TP53	tumour protein p53
TPPI	time proportional phase incrementation
TPR	true positive rate
TSP	(trimethylsilyl)propionate
TTF1	thyroid transcription factor 1
UDP	uridine diphosphate
UMP	uridine monophosphate
UPLC	ultra-performance liquid chromatography
UTP	uridine triphosphate
UV	unit variance
VIP	variable importance in the projection
VLDL	very low-density lipoprotein
WHO	World Health Organisation

1 GENERAL INTRODUCTION

The present chapter starts by briefly describing some general information on the epidemiology, pathogenesis, classification, diagnosis and treatment of lung cancer. Then, current knowledge on cancer cell metabolic reprogramming is presented, addressing also how this information has been exploited in cancer diagnosis and treatment. Thirdly, the metabolomics approach is described, addressing the principles of the analytical and statistical methods employed and presenting a short state of the art of its application in clinical oncology, particularly in the study of lung cancer. In the last section, the scope and aims of this thesis are presented.

1.1 About lung cancer

1.1.1 Epidemiology, aetiology and pathogenesis

The global burden of cancer has doubled in the last third of the twentieth century, and, in 2008, there were 12.7 million cases and 7.6 million deaths worldwide (Ferlay et al. 2010). Based on the continued growth and ageing of the world's population and on the increasing incidence in low- and middle-income countries, this burden could achieve, by 2030, 27 million incident cases and 17 million deaths per year, (Boyle and Levin 2009).

According to the GLOBOCAN database, and as shown in Figure 1.1, lung cancer is one of the cancer types with highest incidence and mortality in the world, being the first cause of cancer-related deaths in men (22.5%) and the second in women (13.8%) (Ferlay et al. 2010). In Portugal, the estimated incidence of lung cancer was (in 2008) 7.6% of all tumours, after colorectal, breast and prostate cancers, while mortality reached 13.7% of cancer-related deaths, when considering both genders (Ferlay et al. 2010). Regarding lung cancer survival rates, the overall 5-year prevalence (percentage of survivors five years after diagnosis) was only 5.8% worldwide and 3.2% in Portugal (Bray et al. 2013), as illustrated in Figure 1.1. Diagnosis at advanced disease stage largely accounts for these numbers, as they significantly improve for regional (25%) and localized disease (52%) (Ridge et al. 2013). Unfortunately, due to the asymptomatic development of lung tumours and the lack of appropriate screening tools, lung cancer diagnosis usually occurs at late stages of tumour progress, resulting in a poor prognosis for the patient.

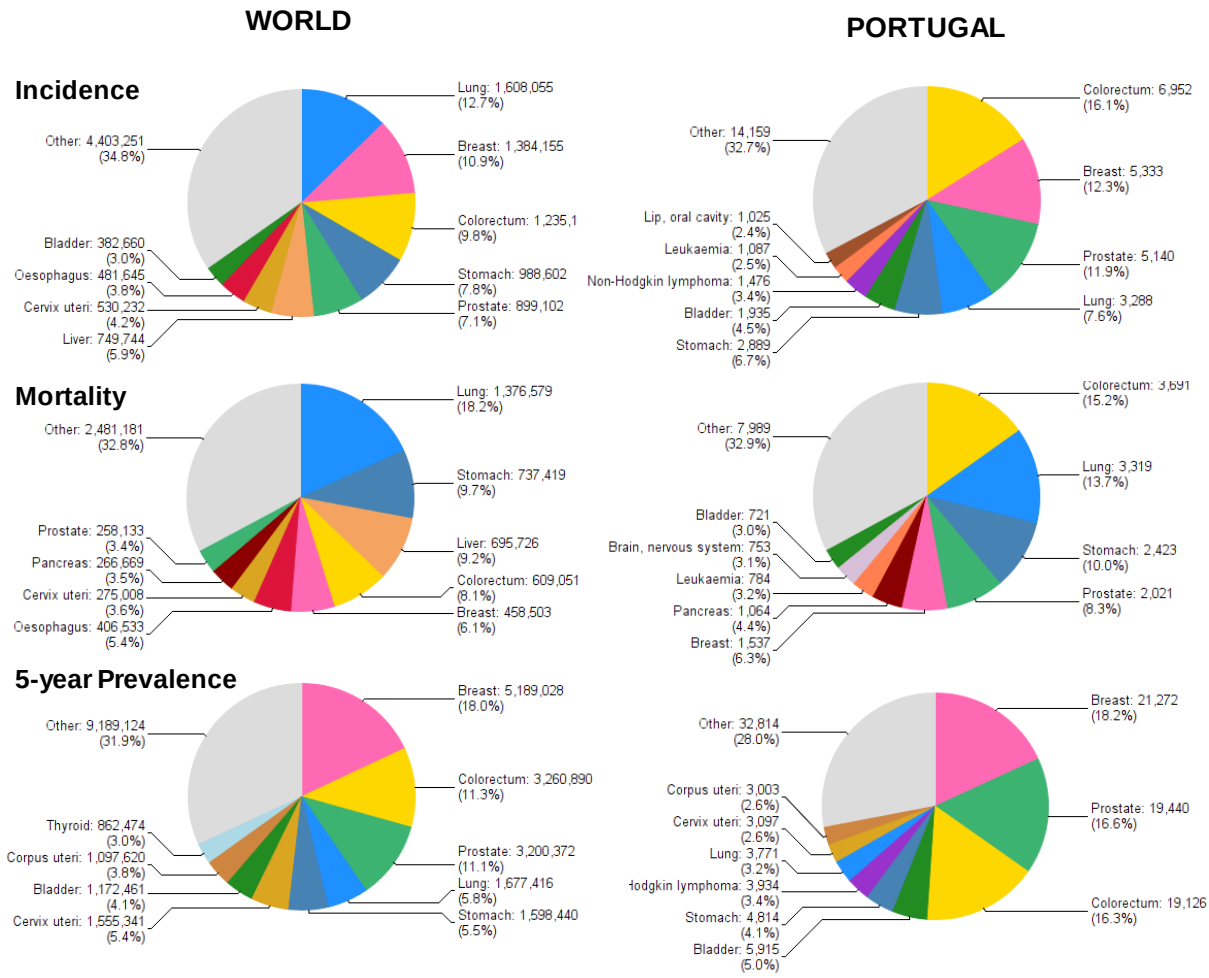


Figure 1.1 Statistical data of cancer incidence, mortality and five-year survival rates worldwide (left) and in Portugal (right). (Adapted from Ferlay et al. 2010).

It is estimated that about 89% of lung cancers in men and 72% in women are caused by carcinogens present in tobacco smoke (e.g., polycyclic aromatic hydrocarbons and tobacco-specific nitrosamines) (Hecht 2012). Concordantly, lung cancer risk increases proportionally with smoke exposure (Boyle and Levin 2009), and global trends in lung cancer incidence and mortality are, to a certain extent, a reflection of population changes in smoking habits (including dose, duration and type of tobacco used). Indeed, age-standardised incidence and mortality rates are, on average, higher in developed regions than in developing countries, reflecting the longer-term smoking habits in the former, while the recent smoking epidemic in medium- and low-resource countries is expected to result, in the future, in a greater number of cancers in those regions (Boyle and Levin 2009). Also, different worldwide trends of tobacco smoking for men and women (smoking prevalence peaked about two decades earlier in men than in women, Alberg et al. 2013)

explains why the mortality rate for men began to decline in the early 90s, while for women it started to decrease later, in 2003 (Ridge et al. 2013). Another interesting relationship with smoking habits regards the shift in the prevalence of different histological types, observed in developed countries over the past 50 years. While adenocarcinoma has become more common, squamous cell carcinoma, which is strongly related to tobacco smoking, has declined, probably as a reflection of changes in cigarette design and composition, as well as in cigarette smoke inhalation patterns (Khuder 2001; Pesch et al. 2012).

In spite of tobacco smoking being, by far, the major aetiological agent for lung cancer, a considerable number of new cases develop in never smokers (25% of all cancer cases, Thun et al. 2008), suggesting other factors to play a role in cancer development. Such factors include family history/heritable factors, exposure to second-hand smoke, occupational and environmental carcinogens (asbestos, polycyclic hydrocarbons, heavy metals and silica, arsenic and chromium VI compounds, diesel engine exhaust), hormonal factors, pre-existing lung disease (chronic obstructive pulmonary disease – COPD, pneumonia, tuberculosis), human immunodeficiency virus infection, human papilloma virus infection, and exposure to ionizing radiation (Alberg et al. 2013; Ridge et al. 2013). The synergistic interaction between exposure to these aetiological agents and individual susceptibility to their effects is at the origin of lung cancer. Nevertheless, reduction of tobacco smoking remains the key strategy for the prevention of lung cancer (Boyle and Levin 2009).

As for cancer in general, lung cancer is characterized by a multitude of genetic and epigenetic alterations, through which normal cells are progressively transformed into tumour cells undergoing uncontrolled proliferation (Weinberg 2013). Among the many genetic alterations associated with lung cancer, the more frequently observed changes include several chromosomal aberrations (aneuploidy, specific allelic loss or gain), activation of key oncogenes (resulting in persistent upregulation of mitogenic growth signals which induce cell growth), and inactivation of tumour suppressor genes (Larsen and Minna 2011). One of such genetic abnormalities present in approximately 70% of patients with non-small cell lung cancer (NSCLC) is the overexpression of the epidermal growth factor receptor (EGFR), a tyrosine kinase (TK) receptor (Franklin et al. 2002). EGFR signalling activates two major pathways in solid tumours, the RAS/RAF/MEK/MAPK pathway and the phosphatidylinositol 3-kinase

(PI3K)/AKT/mTOR pathway, which collectively promote cancer cell growth, proliferation, invasion, metastatic spread, apoptosis and tumour angiogenesis (Reungwetwattana et al. 2012). EGFR tyrosine kinase inhibitors (TKIs), like gefitinib and erlotinib, are important examples of targeted therapy used in the treatment of this type of lung cancer. Another frequently mutated gene in lung cancer is the *KRAS* gene, found in 10-15% of NSCLC, especially in adenocarcinomas (20-30%) (Sekido et al. 2003). *KRAS* mutations result in activation of downstream signalling pathways, such as the PI3K and MAPK, rendering *KRAS* mutant tumours independent of EGFR signalling and therefore resistant to EGFR TKIs (Larsen and Minna 2011). Other commonly activated oncogenes include ERBB2, MYC, MET, CCND1, CDK4, EML4-ALK fusion and BCL2, as reviewed in Larsen and Minna 2011.

Alterations on growth-inhibitory tumour suppressor pathways are also present during lung carcinogenesis, the most important being the p53 pathway (Sato et al. 2007), which is inactivated in ca. 90% of SCLC and 50% of NSCLC (Takahashi et al. 1989). Other recurrently inactivated tumour suppressor genes in lung cancer include RB1, STK11, CDKN2A and FHIT (Larsen and Minna 2011). Most inactivating mutations are point mutations on the DNA binding domain, hence reinforcing the correlation between such mutations and DNA adduct formation by smoke carcinogens (Wang et al. 1995).

1.1.2 Screening and diagnosis

Based on its high incidence and mortality, strongly related to late detection, lung cancer would constitute a good candidate for screening within the general population or high-risk groups. However, randomised, controlled trials performed during the 1970s and 1980s, based on the use of periodical chest X-rays and/or sputum cytology, did not validate the principle that the detection of localised and resectable tumours would greatly improve patients' prognosis (results reviewed by Bach 2003; Aberle et al. 2011; Xiang et al. 2013). Indeed, while more early-stage cancers were detected, mortality rates did not improve and there was no evidence that the tumours found through screening would have developed to an advanced stage (Bach et al. 2007). More recently, in 2011 and 2013, a trial conducted by the National Lung Screening Trial (NLST) used low-dose computed tomography (LDCT), a more sensitive screening modality, together with chest radiography, to detect lung tumours in current or former heavy smokers (Aberle, Berg, et al. 2011). The results

showed a 20% reduction in lung cancer-related deaths within this high-risk group. Despite the promising outcome, this screening tool is not yet ready for a wide population-based implementation, as several issues remain to be addressed: definition of the at-risk population; timing, interval and method of computed tomography; how to handle false positives; the potential toxicity from radiation exposure; and, particularly, the cost-effectiveness when compared with, for example, smoking cessation alone (Aberle et al. 2013). In addition, other screening approaches for the early detection of lung cancer are currently under investigation, such as the search for molecular biomarkers of the disease. Some of the most extensively studied examples are free circulating DNA and RNA, exosomal mRNA, circulating tumour cells, and various lung cancer specific antigens (carcinoembryonic antigen – CEA, serum cytokeratin 19 fragment – CYFRA, and progastrin-releasing peptide - ProGRP), as it has been recently reviewed (Hassanein et al. 2012; Xiang et al. 2013; Brothers et al. 2013).

Lung cancer accurate diagnosis is of paramount importance for establishing the patients' treatment and prognosis. Clinical practice guidelines concerning diagnosis and management of lung cancer were recently revised by the American College of Chest Physicians (ACCP) (Detterbeck et al. 2013) and the European Society for Medical Oncology (ESMO) (Vansteenkiste et al. 2013). However, the definition of the diagnostic workflow is usually decided on a case-by-case basis, largely depending on the overall clinical evaluation of the patient (including physical examination, medical history and performance status), the size and location of the tumour and the presumed stage of disease (Detterbeck et al. 2013; Gould et al. 2013).

Available diagnostic tools typically comprise imaging techniques, laboratory testing and tissue microscopic examination after collection of a biopsy. The overall sensitivity and specificity of some of these tools are summarised in Table 1.1. Imaging methods like computed tomography (CT) scans are useful to determine the size and location of the tumours and to detect affected mediastinal/regional lymph nodes, while positron emission tomography (PET) scans may further assist in the differentiation of benign and malignant tumours (based on their metabolic activity) and in detecting metastasis to distant sites. Laboratory testing, such as sputum cytology and measurement of circulating tumour markers, may be performed when lung cancer is suspected, although poor sensitivity and/or specificity are usually achieved. Examination of a tissue sample under the

microscope is crucial for a definite diagnosis. Several techniques may be used for collecting a tissue biopsy, such as flexible bronchoscopy (FB, especially useful for biopsying tumours in the central part of the lung), transthoracic needle aspiration (TNA, recommended for peripheral lesions) and mediastinoscopy (Detterbeck et al. 2013). Subsequent histopathological study of tissue biopsies comprises their morphological assessment by light microscopy and, eventually, immunohistochemical testing, which is especially useful in cases of poorly differentiated tumours or small biopsy specimens.

Table 1.1 Sensitivity and specificity of several lung cancer diagnostic tools.

Diagnostic tools	**Sensitivity (%)**	**Specificity (%)**
CT[a]	60	77
FDG-PET[b]	97	78
Sputum cytology[c]	66	99
Pleural fluid cytology[d]	72	---
CEA, in blood[e]	69	68
CYFRA 21-1, in blood[e]	43	89
Flexible bronchoscopy (FB), for central endobronchial lesions[c]	88	---
Flexible bronchoscopy (FB), for peripheral lesions[c]	34-63	---
Transthoracic needle aspiration (TNA) or biopsy[c]	90	---

[a] Dwamena et al. 1999; [b] Gould et al. 2001; [c] Rivera and Mehta 2007; [d] Rivera et al. 2013; [e] Okamura et al. 2013.

1.1.3 Main histological types of lung tumours

According to the World Health Organization (WHO) (Travis et al. 2004, Travis et al. 2013), malignant epithelial lung tumours are classified into seven main categories (as listed in Table A1 of Annex I): adenocarcinoma, squamous cell carcinoma, adenosquamous carcinoma, sarcomatoid carcinoma, carcinoid tumour, large cell carcinoma and small cell carcinoma, the first six categories traditionally being grouped as a larger class designated as non-small cell lung carcinoma (NSCLC). Accurate tumour subtyping is a major requirement as several clinical trials have demonstrated differing efficacy and/or toxicity of particular treatments depending on tumour histological type (Cooper et al. 2011; Kerr 2012). This task is typically based on stained hematoxylin and eosin (H&E) histological sections, as well as on immunohistochemical assays. The following sub-sections describe

in more detail the main histological, cytological and clinical features of the tumour types included in this work.

1.1.3.1 Adenocarcinoma (AdC)

Adenocarcinoma is a malignant epithelial tumour with glandular differentiation or mucin production, showing one or a mixture of the following histological patterns: acinar, papillary, bronchioloalveolar (BAC), solid with mucin production (Travis et al. 2004). Adenocarcinomas of mixed patterns are by far the most frequent subtype, representing ca. 80% of resected adenocarcinomas (Terasaki et al. 2003), and often presenting several degrees of differentiation. Typical cytomorphological features of AdC include three-dimensional cellular clusters of large vacuolated cells, with variable cell size and cytoplasm volume (Figure 1.2) (Travis et al. 2004; Idowu and Powers 2010). The nuclei are usually single, eccentric and round to oval. The chromatin tends to be finely regular and evenly dispersed in well-differentiated tumours, while coarse and irregularly distributed in poorly differentiated tumours (Travis et al. 2004). Adenocarcinoma spreads primarily by lymphatic and haematogenous routes, with nearly one fifth of the newly diagnosed adenocarcinomas presenting distant metastasis, mainly to the brain, bone, adrenal glands or liver (Quint et al. 1996). The immunohistochemical features of AdCs may vary according to the subtype and degree of differentiation. The expression of epithelial markers (cytokeratins AE1/AE3, CAM 5.2 and CK7, epithelial membrane antigen – EMA, and carcinoembryonic antigen – CEA) is typical of this type of tumour, as well as the presence of thyroid transcription factor 1 (TTF1) staining, in particular in better-differentiated tumours (Rubin et al. 2001; Lau et al. 2002; Yatabe et al. 2002). Nevertheless, mucinous tumours may represent exceptions to these markers. In terms of incidence, adenocarcinoma has surpassed squamous cell carcinoma as the commonest histological subtype of lung cancer, largely associated with the introduction of filter cigarettes (Ito et al. 2011). Incidence rates go up to 41.5 % and 39.6% in the USA and Portugal, respectively (Hespanhol et al. 2013; Howlader et al. 2014).

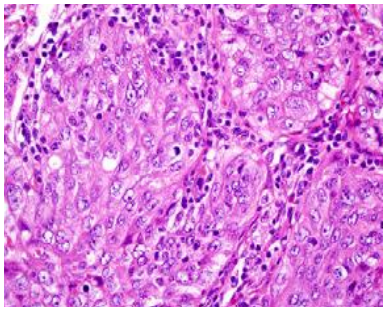


Figure 1.2 Solid adenocarcinoma with mucin (hematoxylin-eosin). (Adapted from Travis et al. 2004).

1.1.3.2 Squamous cell carcinoma (SqCC)

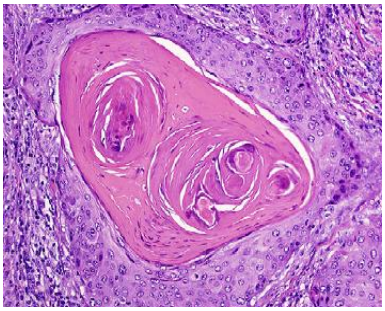


Figure 1.3 Squamous cell carcinoma (hematoxylin-eosin). (Adapted from Travis et al. 2004).

Squamous cell carcinoma, also referred to as epidermoid carcinoma, is a type of malignant epithelial tumour, usually located centrally in the mainstem, lobar or segmental bronchi (Tomashefski et al. 1990), although an increase of the frequency of peripheral carcinomas has been noted (Funai et al. 2003). The variants of this histological type are: papillary, clear cell, small cell and basaloid carcinomas. Well-differentiated squamous cell carcinomas show keratinisation and/or intercellular bridges that arise from bronchial epithelium and pearl formation (Figure 1.3). Tumour cells tend to be large with abundant dense cytoplasm, central and irregular hyperchromatic nuclei and small nucleoli (Travis et al. 2004). In the case of poorly-differentiated tumours, these features can be present, although only focally. Metastases to distant organs are less frequent than in other histological types of primary lung cancer (Quint et al. 1996). The majority of squamous cell carcinomas consistently express high-molecular weight keratin (34βE12), cytokeratin CK5/6 and CEA. In addition, these carcinomas frequently express protein P63 and are negative for TTF1 (Travis et al. 2004; Drilon et al. 2012). Over 90% of cases with SqCC occur in cigarette smokers (Spiro and Porter 2002). Being the second most common type of lung cancer, incidence rates reach 22.0% and 26.8 % in the USA and Portugal, respectively (Hespanhol et al. 2013; Howlader et al. 2014).

1.1.3.3 Adenosquamous carcinoma (ASqCC)

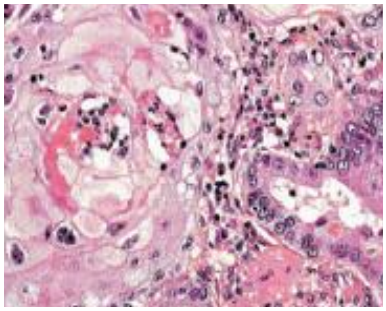


Figure 1.4. Adenosquamous carcinoma, showing a squamous cell lobe on the left, adjacent to an acinar adenocarcinoma. (Adapted from Travis et al. 2004).

Adenosquamous carcinoma is characterised by showing components of both AdC and SqCC with each comprising at least 10% of the tumour, as identified by light microscopy, thus resection specimens are required for diagnosis. This hybrid type of tumour accounts for 0.6-2.3% of all lung cancers worldwide (Takamori et al. 1991; Sridhar et al. 1992; Ishida et al. 1992) and for 0.9% in Portugal (Hespanhol et al. 2013). ASqCCs are usually located in the periphery of the lung and may contain a

central scar. Immunohistochemical findings also recapitulate both squamous and adenocarcinoma characteristics, so they express cytokeratins with a wide molecular weight range including AE1/AE3, CAM 5.2, KL1, and CK7, but usually not CK20. EMA is positive and TTF1 positivity is confined to the adenonocarcinoma component (Travis et al. 2004).

1.1.3.4 Sarcomatoid carcinoma (SC)

Sarcomatoid carcinoma is a poorly differentiated carcinoma which contains a component of sarcoma or sarcoma-like differentiation. Five subtypes can be recognised within sarcomatoids: pleomorphic, spindle cell, giant cell carcinomas, carcinosarcoma and pulmonary blastoma (Travis et al. 2004). SCs can arise in the central or peripheral lung, accounting for only 0.3-1.3% of all lung malignancies (Fishback et al. 1994; Nakajima et al. 1999; Rossi et al. 2003; Travis 2010a). Pleomorphic carcinomas tend to be large, peripheral tumours that can invade the chest wall, and are often associated with poor prognosis (Travis 2010a). They consist of malignant giant and/or spindle cells (at least 10%) and epithelial components, such as squamous or adenocarcinoma (Figure 1.5) (Travis et al. 2004; Travis 2010a), their diagnosis usually requiring a resection specimen. On the other hand, the spindle and/or giant cells can occur as cohesive aggregates, lacking any glandular or squamous differentiation, forming the spindle cell or the giant cell carcinoma, respectively. Sarcomatoid cells often co-express cytokeratins, vimentin, CEA and smooth markers (Addis et al. 1988; Chejfec et al. 1991; Attanoos et al. 1998; Fishback et al. 1994; Rossi et al. 2003) and TTF1 may be positive in giant cell carcinomas.

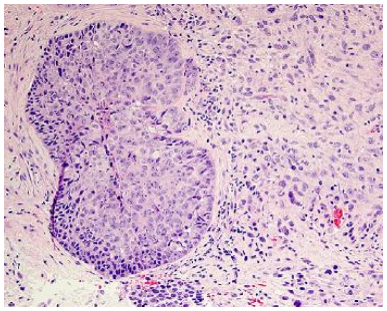


Figure 1.5. Pleomorphic carcinoma, composed of squamous cell carcinoma (left) and a malignant spindle cell proliferation (right) (hemato-xylin-eosin). (Adapted from Travis 2011).

1.1.3.5 Carcinoid tumour

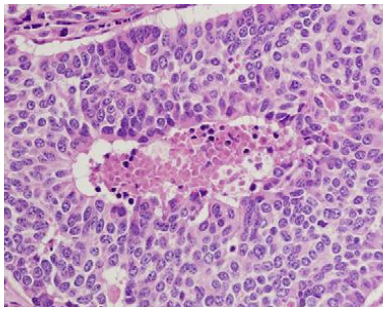


Figure 1.6. Atypical carcinoid with foci of necrosis in the center of several organoid nests of uniform tumour cells. (Adapted from Travis 2011).

Carcinoid tumour cells are characterised by growth patterns (organoid, trabecular, insular, palisading, ribbon, rosette-like arrangements) that suggest a neuroendecrine differentiation (Travis et al. 2004), with nuclei posessing finely granular chromatin in eosinophilic cytoplasm. They can be classified as typical carcinoid (tumour with fewer than 2 mitoses per 2 mm^2 and lacking necrosis) and atypical carcinoid (tumour with 2-10 mitoses per 2 mm^2 and/or focal presence of necrosis, still below the mitotic rates of large cell neuroendocrine and small cell lung carcinomas) (Travis et al. 1991; Travis et al. 1998). Carcinoids account for 1-2% of all invasive lung malignancies (Travis et al. 1995), 50% of patients being asymptomatic at presentation (Travis et al. 1995; Asamura et al. 2006; Travis 2010b). The 5-year survival rates of typical carcinoid (90-95%) largely surpass the ones of atypical carcinoid (50-60%) (Travis 2011). Carcinoid tumours stain for neuroendocrine markers, such as chromogranin, synaptophysin and CD56 (neural cell adhesion molecule – NCAM) (Travis et al. 2004).

1.1.3.6 Large cell carcinoma (LCC)

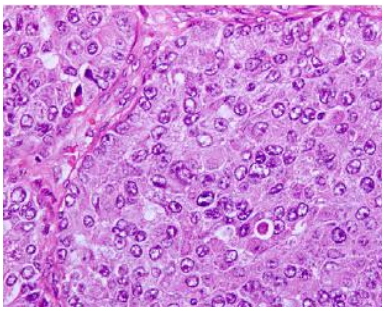


Figure 1.7 Large cell carcinoma (hematoxylin-eosin). (Adapted from Travis et al. 2004).

Large cell carcinoma is an undifferentiated cell carcinoma without cytological and architectural features of neither small cell carcinoma nor glandular or squamous differentiation, so it is diagnosed by exclusion (Travis et al. 2004). Large cell neuroendocrine carcinoma, combined large cell neuroendocrine carcinoma, basaloid carcinoma, lymphoepithelioma-like carcinoma, clear cell carcinoma and large cell carcinoma with rhabdoid phenotype comprise the list of this tumour variants. This type of lung tumour accounts for approximately 3% of all lung carcinomas (Iyoda et al. 2001; Hespanhol et al. 2013) and typically presents as large, peripheral masses (although central location may also be found). Cell nuclei vary from round to extremely irregular, with irregular chromatin

distribution (Burns et al. 1989). Patients with this type of lung cancer have a poor prognosis, worse than patients with other carcinomas, in particular adenocarcinomas (Travis et al. 1998; Jiang et al. 1998; Iyoda et al. 2001).

1.1.3.7 Small cell carcinoma (SCLC)

Small cell carcinoma of the lung is a malignant epithelial tumour comprised of small cells with a round to fusiform shape, scant cytoplasm, ill-defined cell borders, finely granular and uniformly distributed nuclear chromatin ('salt and pepper' effect) and absent or inconspicuous nucleoli. Tissue necrosis is typically extensive and the mitotic count is high (average rate of 80 mitoses/2 mm^2) (Nicholson et al. 2002; Travis et al. 2004; Travis 2010b). Two variants are small cell lung cancer and combined small cell lung cancer (with an additional component of any of the histological types of non-small cell carcinoma). Immunohistochemistry is positive for neuroendocrine markers CD56, chromogranin and synaptophysin in most cases (Guinee et al. 1994; Nicholson et al. 2002), and for TTF1 in up to 90% of the cases (Folpe et al. 1999; Kaufman and Dietel 2000). SCLC comprises up to 13% of all lung cancers (Travis et al. 1995; Hespanhol et al. 2013; Howlader et al. 2014) and often presents an aggressive clinical course, with tendency to metastasise (widespread at presentation and poor responsiveness to chemotherapy).

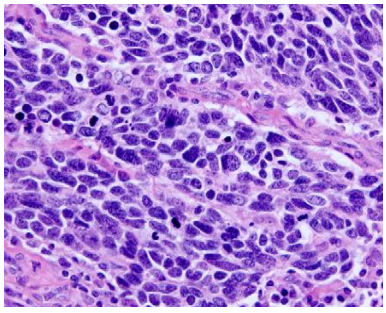


Figure 1.8. Small cell carcinoma. (Adapted from Travis 2011).

1.1.4 Staging and treatment

Lung cancer staging is based on the internationally accepted tumour-node-metastasis (TNM) staging system, which assesses the extent of the primary tumour (T), the absence or presence of regional lymph node involvement (N), and the absence or presence of intrathoracic or distant metastasis (M) (Lababede et al. 2011; Mirsadraee et al. 2012). Disease stages range from I (less advanced) to IV (more advanced), each comprising different TNM classifications (Tables A2 and A3 of Annex II).

The treatment strategy is largely defined based on tumour histology and stage, other factors taken into account include the patients' age, performance status, comorbidities and preferences. The primary curative approach to early-stage NSCLC (I and II) is usually surgery, either by removing part of the lung (lobectomy) or all of the lung (pneumonectomy), depending on the location of the tumour (Vansteenkiste et al. 2013). If tumour resection is complete, adjuvant chemotherapy is administered to lower the risk of recurrence. In cases of incomplete resection, postoperative radiotherapy is usually offered. For treating locally advanced NSCLC of stage III, surgery may also be considered, preferably in patients in whom resection by lobectomy is expected to be successful. Moreover, for both resectable and unresectable tumours, regimens of chemotherapy (mostly cisplatin-based), delivered concurrently with radiotherapy, are recommended (Vansteenkiste et al. 2013). As for stage IV NSCLC patients, chemotherapy is often used to control the disease for as long as possible, while radiation may be used for palliative care. Among the drugs most commonly used for NSCLC treatment are cisplatin or carboplatin, plus docetaxel, gemcitabine, paclitaxel, vinorelbine or pemetrexed (Vansteenkiste et al. 2013; Peters et al. 2012). In cases of advanced tumours with specific molecular markers, targeted therapies may also be employed, alone or in combination with conventional chemotherapy (e.g., erlotinib to treat EGFR mutated tumours).

Recommended treatment regimens for SCLC differ significantly, as this type of cancer is characterised by rapid growth, high response rates to both chemo and radiotherapy, and development of treatment resistance in patients with metastatic disease (Früh et al. 2013). In this case, surgery may benefit a small percentage of patients (5% of SCLC patients with stage I localized disease), while chemotherapy is essential at all stages. Patients with limited-stage SCLC are often treated with concurrent chemotherapy (using drug combinations like etoposide and cisplatin) and thoracic radiotherapy. Treatment of advanced SCLC is mostly palliative, while in patients with relapsed or refractory SCLC (patients not responding or progressing during chemotherapy), the administration of second-line, single agent chemotherapy, or participation in a clinical trial is recommended (Früh et al. 2013; Detterbeck et al. 2013).

1.2 Metabolic reprogramming in cancer

Tumour development involves a complex network of events, such as oncogene activation, insensitivity to anticancer signalling and evasion to apoptosis, high replicative potential, sustained angiogenesis and metastasis, and metabolic dysregulation (Hanahan and Weinberg 2000; Hanahan and Weinberg 2011). In particular, altered metabolism of cancer cells has been recently recognised as an emergent cancer hallmark, resulting from changes in signalling pathways, protein expression and other molecular mechanisms, but also reflecting specific biochemical adaptations during carcinogenesis, with extensive cross-talk between events, as well as important feedback loops (Hanahan and Weinberg 2011). This means that the metabolic status of tumour cells will reflect not only the endpoint of oncogenic activation, but may also actively participate in the process of tumour development by conferring malignant cells survival advantages. Thus, an accurate mapping of tumour metabolic signatures may potentially provide new insights into tumour cell biology and offer new options for diagnosing and treating cancer.

1.2.1 Altered metabolic pathways in cancer and their regulation

Over the last decades, studies on diverse types of tumour cells have allowed consistent alterations in multiple metabolic pathways to be identified. In general, metabolism in these cells is characterised by high rates of glycolysis and/or glutaminolysis, to fuel cell proliferation, and increased biosynthetic processes to provide essential building blocks (nucleotides, lipids and other macromolecules), as well as reducing power (NADPH) for redox regulation. The following sections will summarily describe the most widely recognised metabolic alterations in cancer, addressing also their regulation mechanisms. An overview representation of cancer metabolic pathways and the key oncogenes and tumour suppressors involved in their regulation is shown in Figure 1.9.

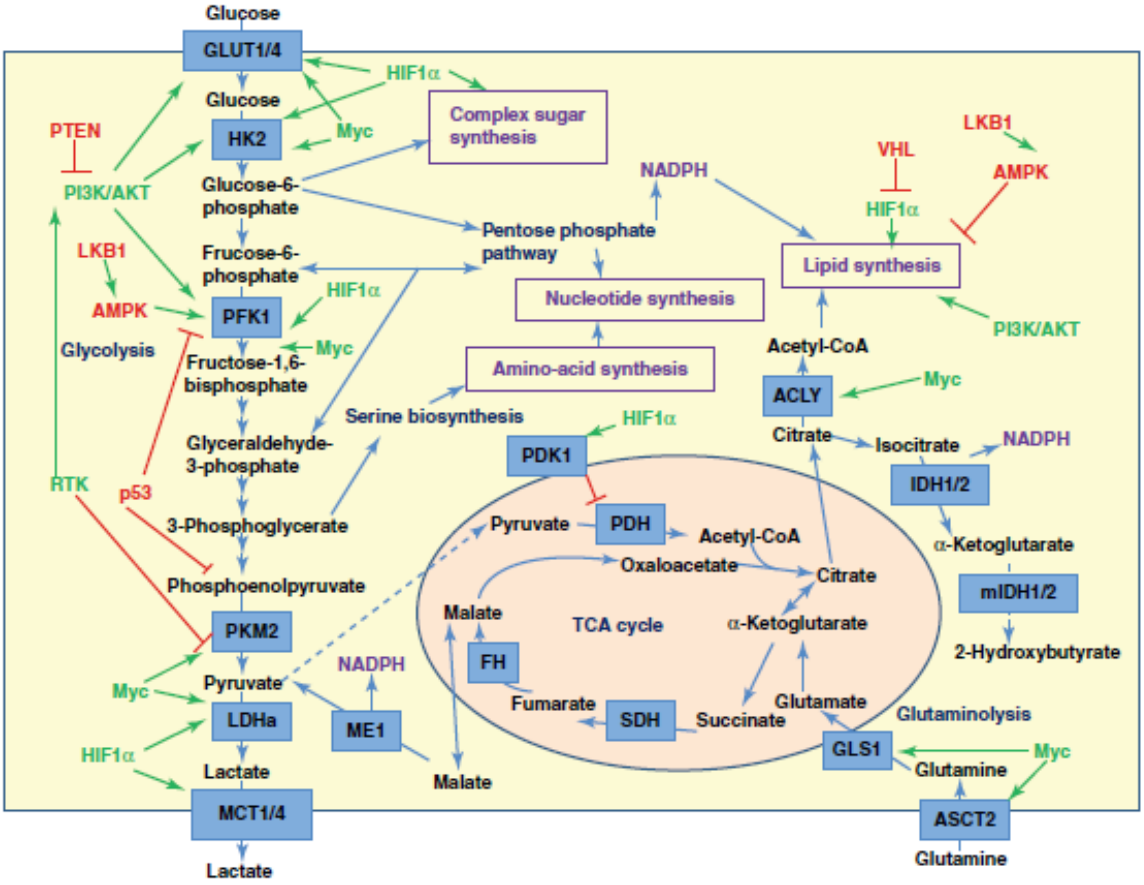


Figure 1.9 Overview of altered cancer metabolic pathways and their regulation. Key oncogenic pathways are shown in green and key tumour suppressor pathways are shown in red. (Reprinted from Jones and Schulze 2012, Copyright © 2011 with permission from Elsevier).

1.2.1.1 Tumour energy supply: glycolysis and glutaminolysis

In contrast to nonproliferating (differentiated) cells, which depend primarily on oxidative phosphorylation for ATP production, proliferative tissues or tumour cells rely preferentially on glycolysis to satisfy their energetic needs, even in the presence of oxygen (aerobic glycolysis) (Figure 1.10). Glycolysis is the process by which glucose enters the cell and is degraded, via a cascade of reactions in the cytosol, into pyruvate, which can then either be incorporated into the tricarboxylic acid (TCA) cycle or be converted to lactate (Nelson and Cox 2004). The glycolytic degradation into pyruvate releases only a small fraction of the total available energy from glucose. So, under aerobic conditions, this energy can be further extracted by oxidative reactions in the TCA cycle and through oxidative phosphorylation pathway, in the mitochondria. Oppositely, under low oxygen conditions (hypoxia), pyruvate is reduced into lactate by lactate dehydrogenase (LDH). As firstly observed by Otto Warburg in 1924, tumour cells are characterised for uptaking more

glucose than normal cells and for having a higher glycolytic capacity and higher rate of lactate production, even in the presence of oxygen (Warburg et al. 1924; Warburg 1956); a phenomenon called the 'Warburg effect'.

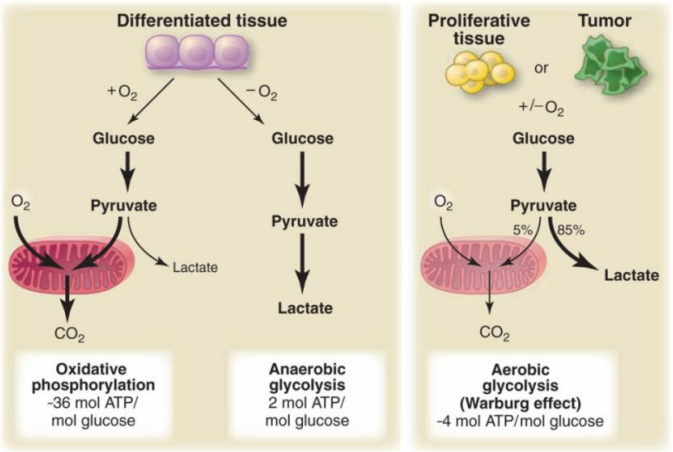


Figure 1.10 Schematic representation of the different utilization of glucose by differentiated and proliferative or tumour tissues. From Heiden et al. 2009, reprinted with permission from AAAs.

Tumour cells commonly experience hypoxia, as they initially lack an extensive capillary network to supply the tumour with oxygen. In these conditions, cells adjust their cellular physiology and metabolism, switching from the oxygen-dependent oxidative phosphorylation to the oxygen-independent glycolysis. In mammalian cells, this metabolic response to oxygen limitation is coordinated by the transcription factor hypoxia-inducible factor 1α (HIF-1α) encoded by the *HIF-1A* gene (Semenza 2003; Gordan and Simon 2007). The activation of *HIF-1A* gene in cancer cells drives the overexpression and increased activity of several glycolysis-related protein isoforms that differ from those found in non-malignant cells, including hexokinase HK2 and phosphofructokinase PFK1, as well as of glucose transporters (GLUT1), pyruvate dehydrogenase kinase (PDK1), which blocks the entry of pyruvate into the TCA, and lactate dehydrogenase LDHa (Macheda et al. 2005; Kim et al. 2006; Marín-Hernández et al. 2009) (Figure 1.9).

However, hypoxia cannot be completely responsible for the elevated glucose transport and increased glycolysis, as these alterations have also been verified in cultured cells under normoxic conditions. In fact, recent studies demonstrated that besides the referred HIF-induced alterations in the expression of key glycolytic enzymes, other changes contribute to the 'Warburg effect' in cancer cells, namely the inactivation of *p53*,

a tumour suppressor gene that promotes oxidative phosphorylation and inhibits glycolysis (Gottlieb and Vousden 2010), and the activation of several oncogenes, including *MYC*, *RAS*, *SRC* and *AKT* (Dang and Semenza 1999; Bensaad and Vousden 2007; Young and Anderson 2008). For instance, the activation of the PI3K/AKT pathway and MYC can result in the increase of glucose uptake and activation of several glycolytic enzymes (Elstrom et al. 2004; Yeung et al. 2008), as shown in Figure 1.9.

Although glycolysis appears to be an inefficient way of energy production (low ATP yield per glucose molecule consumed), it provides several advantages for tumour cells. First, glycolytic ATP production is independent of oxygen availability and, if the glycolytic flux is high enough, it can be a faster way of obtaining energy than oxidative phosphorylation. Second, glycolysis provides cells with metabolic intermediates necessary for biosynthetic processes (as described in the next section), the high glycolytic rate allowing those lower-flux pathways to be finely controlled. Thirdly, the high rate of lactate production from glycolysis results in an acidic, toxic microenvironment that can actually favour tumour cells as, contrarily to other local cell populations, these cells tend to develop a phenotype resistant to hypoxia and to acid-induced cell death (necrosis or apoptosis). Ultimately, this promotes extracellular matrix degradation and aids invasiveness and metastasis (Fang et al. 2008).

In addition to glycolysis, glutaminolysis is also recognised as a key feature of cancer cells' metabolism. Glutamine, the most abundant free amino acid in plasma, serves as a carbon source for energy production, contributes with carbon and nitrogen to biosynthetic reactions, and helps regulating the redox status (Daye and Wellen 2012). It is known, since decades, that the high consumption rate of glutamine observed in cancer cells largely exceeds the one necessary for amino acid and protein synthesis (Eagle et al. 1956). Later studies showed glutamine to be partially oxidised into lactate, and established glutamine as an additional source of energy in tumour cells (Reitzer et al. 1979). This catabolic degradation of glutamine into lactate is termed glutaminolysis, another hallmark in cancer cell metabolism.

Glutaminolysis starts with transport of glutamine into the cell by transporters ASCT2 and SN2, and proceeds with glutamine deamidation to glutamate via glutaminase (GLS), yielding ammonium. Glutamate can be excreted or further metabolised to α-ketoglutarate via transaminase alanine aminotransferase (ALT). In the mitochondria, α-ketoglutarate is

used in the production of malate, which can then be oxidized to pyruvate via malic enzyme (ME), ultimately leading to the secretion of lactate into the extracellular medium (Wise et al. 2008) (Figure 1.11).

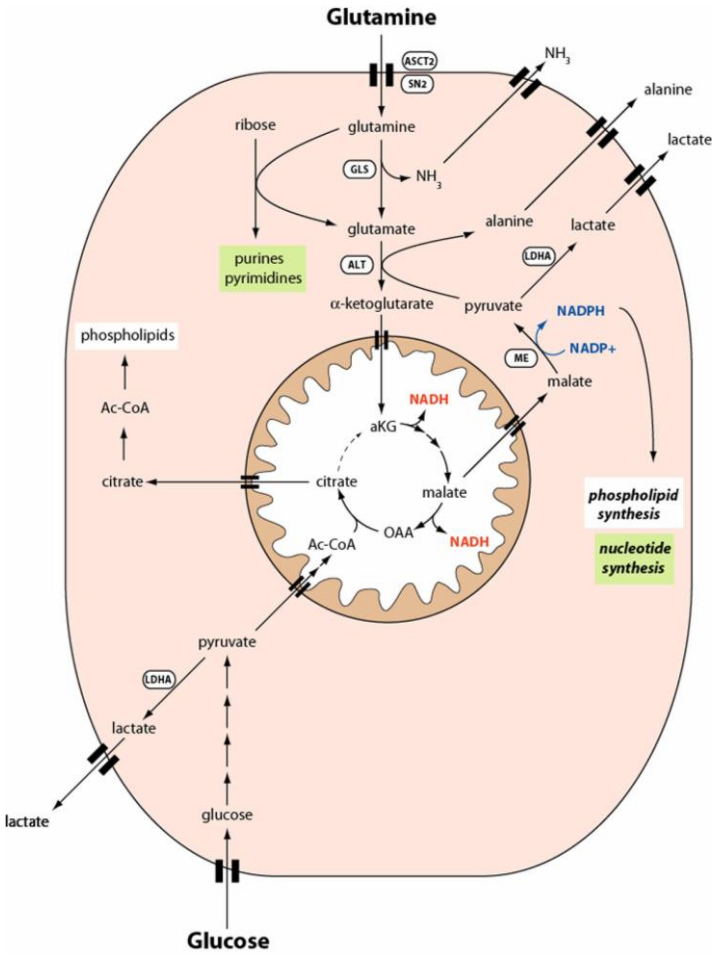


Figure 1.11 Glutaminolytic and glycolytic pathways showing the synthetic endpoints of their metabolic intermediates. (Reprinted from Wise et al. 2008, Copyright © 2008 by The National Academy of Sciences of the USA.).

Although little is known about the regulation of glutamine uptake and of glutaminolysis, recent studies have proposed an important role for the oncogene *MYC* (Wise et al. 2008; Gao et al. 2009). In particular, Gao and colleagues have reported that the overexpression of MYC results in a greater expression of mitochondrial glutaminase, leading to the upregulation of mitochondrial glutamine metabolism (Gao et al. 2009), as shown in Figure 1.9. In addition to MYC, tumour suppressor p53 also regulates glutamine metabolism by inducing the expression of GLS2, leading to increased mitochondrial oxidative phosphorylation and energy production from glutaminolysis (Daye and Wellen

2012). Also, GLS2 was found to increase glutathione levels and to reduce ROS levels, conveying protection against oxidative stress-induced apoptosis (Hu et al. 2010).

Together, glucose and glutamine serve as the primary nutrients to fuel cancer cell proliferation. Recent data indicate that the metabolism of these two nutrients is actively coordinated, as cells can switch the carbon source or compensate for reduced availability of one nutrient by utilizing more of the other (Yang et al. 2009; Cheng et al. 2011). However, in contrast to glycolytic ATP production, glutaminolytic ATP generation depends on oxygen supply and takes place mainly in small tumours with good vascularisation and adequate oxygen levels (Vaupel et al. 1989).

1.2.1.2 Tumour cell biosynthetic pathways

Proliferating cells, in particular tumour cells, depend on the availability of metabolic precursors for the synthesis of cellular building blocks needed for maintaining their high proliferative capacity. One paradigmatic example comprises the diversion of glycolytic intermediates into the pentose phosphate pathway (PPP) and the use of nonessential amino acids, derived from glucose and glutamine catabolism, for *de novo* nucleotide biosynthesis (Figure 1.12). Purine and pyrimidine synthesis utilizes ribose 5-phosphate (R5P) that can be either synthesised from glucose-6-phosphate (glucose-6-P), by the oxidative branch of PPP, or from fructose-6-phosphate (fructose-6-P) and glyceraldehyde-3-phosphate (GA3P), by the non-oxidative branch of PPP (Figure 1.12). Along with the production of R5P, the oxidative PPP also regenerates NADPH, which is the hydrogen donor in fatty acid synthesis, as well as an important regenerator of reduced glutathione, promoting the scavenging of reactive oxygen species (ROS) (Wood 1986).

Unlike normal cells, which produce most of the R5P for nucleotide biosynthesis through the oxidative arm of PPP, there is evidence that, in cancer cells, the non-oxidative pathway is the main source for R5P synthesis (Boros et al. 1998; Cascante et al. 2000), as corroborated by the increased activity of transketolase and transaldolase enzymes (Heinrich et al. 1976; Coy et al. 2005). All the reactions in the non-oxidative PPP are reversible, meaning that the direction of this pathway is determined by the relative levels of metabolic substrates and products. Therefore, in order to divert glycolytic metabolites into PPP through the non-oxidative branch, cancer cells have to maintain high levels of fructose 6-phosphate and/or GA3P. The activity of phosphofructokinase 1 (PFK-1), an

important control point in glycolysis, and the levels of fructose 1,6-biphosphate (F1,6BP), which then breaks down into GA3P by aldolase, have been found increased in cancer cells and tissue, and thought to contribute to this redirection (Sanchez-Martinez and Aragon 1997; Tong et al. 2009). Along with PFK-1, pyruvate kinase M2 (PK-M2), another rate-limiting enzyme in glycolytic degradation, also contributes to divert the carbon flux from glycolysis into PPP (Mazurek et al. 2005). Moreover, besides stimulating anaerobic glycolysis like previously explained, HIF-1a also has an important role in this preference for the non-oxidative PPP, namely by inducing the expression of TKT and PK-M2 (Tong et al. 2009). Furthermore, MYC-induced glutamine catabolism provides the cell with an abundant supply of aspartate and amine groups to support nucleotide biosynthesis, together with NADPH. This is especially important for tumour cells relying on the non-oxidative arm of PPP for R5P production, as the activity of glucose-6-phosphate dehydrogenase (G6PD) is inhibited and the oxidative arm of PPP cannot be used to produce NADPH to support macromolecular biosynthesis.

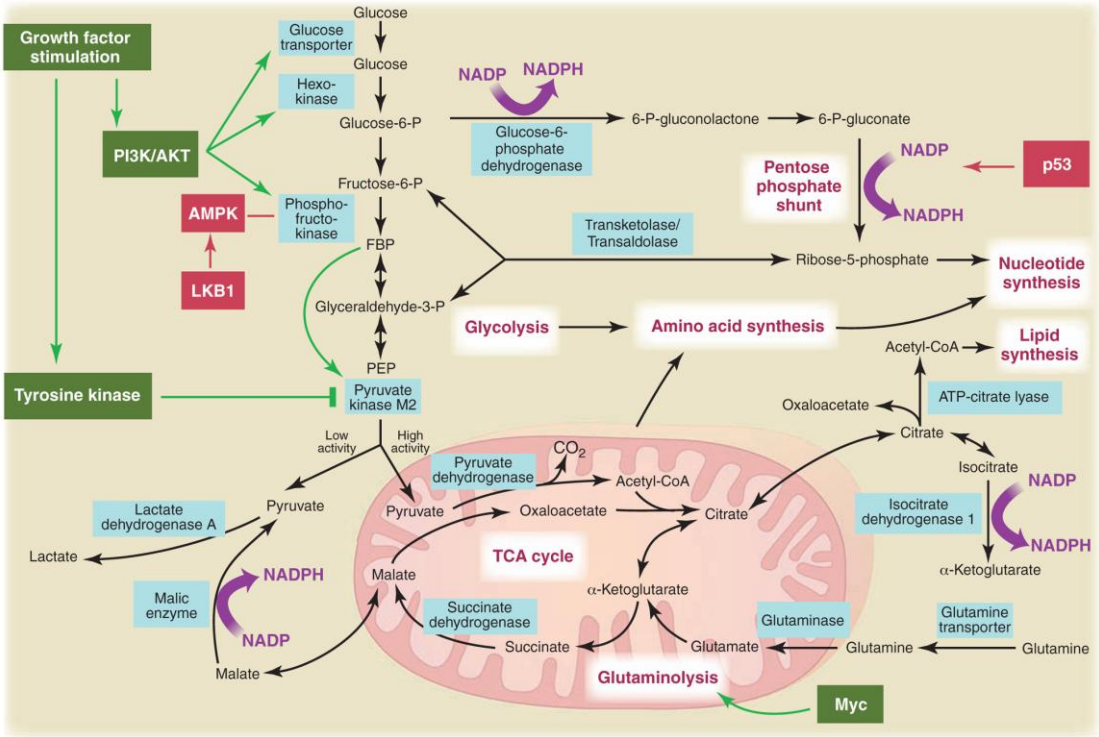


Figure 1.12 Metabolic pathways active in proliferating cells, where the diversion from glycolysis into the pentose phosphate pathway (PPP) is put in evidence. From Heiden et al. 2009, reprinted with permission from AAAs.

Another branch diverting from glycolysis and implicated in cancer is the serine biosynthesis pathway, in which the glycolytic intermediate 3-phosphoglycerate is converted into serine following a three-step enzymatic reaction, accompanied by glutamate breakdown to α–ketoglutarate (Figure 1.13), thereby coupling glycolysis with glutaminolysis. Genetic and functional studies have suggested the serine biosynthetic pathway to be activated in cancer pathogenesis, with p53 having been associated with tumour cells' capacity to deal with serine depletion and oxidative stress (Maddocks et al. 2013). Moreover, serine can act as an allosteric regulator of PKM2, since when serine is abundant this pyruvate kinase is fully activated allowing the consumption of glucose through aerobic glycolysis (Ye et al. 2012; Chaneton et al. 2012). Upon serine deprivation, PKM2 activity is reduced, and pyruvate is directed to mitochondria and glycolysis metabolites are diverted to serine biosynthetic pathways to sustain cell proliferation.

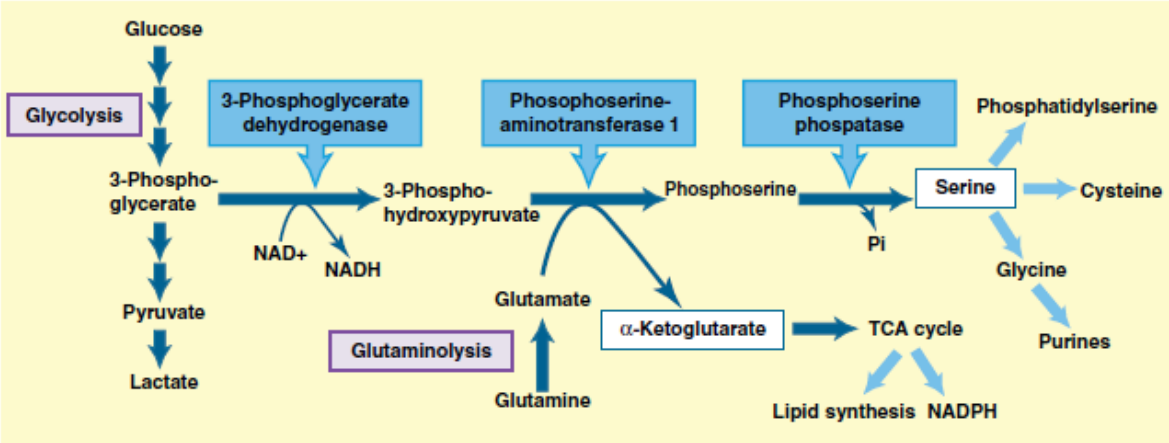


Figure 1.13 Serine biosynthesis pathway. (Reprinted from Jones and Schulze 2012, Copyright © 2011 with permission from Elsevier).

When serine is metabolized to glycine, tetrahydropholate (THF) is converted to 5,10-methylene-THF, which is a critical step in maintaining the folate cycle for nucleotide synthesis. Moreover, glycine can be directed to the synthesis of purines giving two carbon and one nitrogen atoms in the purine ring. Also, glycine is an integral component of glutathione, the main cellular antioxidant. It has been recently suggested a key role for glycine in cancer cell proliferation (Jain et al. 2012), making it another target for therapy intervention.

In addition to glycolysis, the TCA cycle also acts in tumour cells as a hub for biosynthesis, resulting in a continuous efflux of intermediates (cataplerosis). A prime example is the *de novo* synthesis of lipids (fatty acids, cholesterol and isoprenoids), observed in many tumour cells irrespective of the levels of extracellular lipids. In this pathway (Figure 1.14), mitochondrial citrate is exported to the cytosol to be converted into the lipogenic precursor acetyl-CoA. A portion of acetyl-CoA is then carboxylated to malonyl-CoA, and the consecutive condensation of the two coenzymes produces palmitate and other saturated long chain fatty acids (FA), which can be further modified to form more complex FA used in the synthesis of various cellular lipids (phospholipids, triglycerides and cholesterol esters) or in the modification of proteins, thus influencing cell signalling and growth. This increased lipogenic activity impacts on the fraction of mitochondrial citrate available for oxidation, thus resulting in a so called 'truncated' TCA cycle (Mazurek 2007). As already described, tumour cells tend to compensate for this and to sustain the TCA cycle through anaplerotic processes like glutaminolysis.

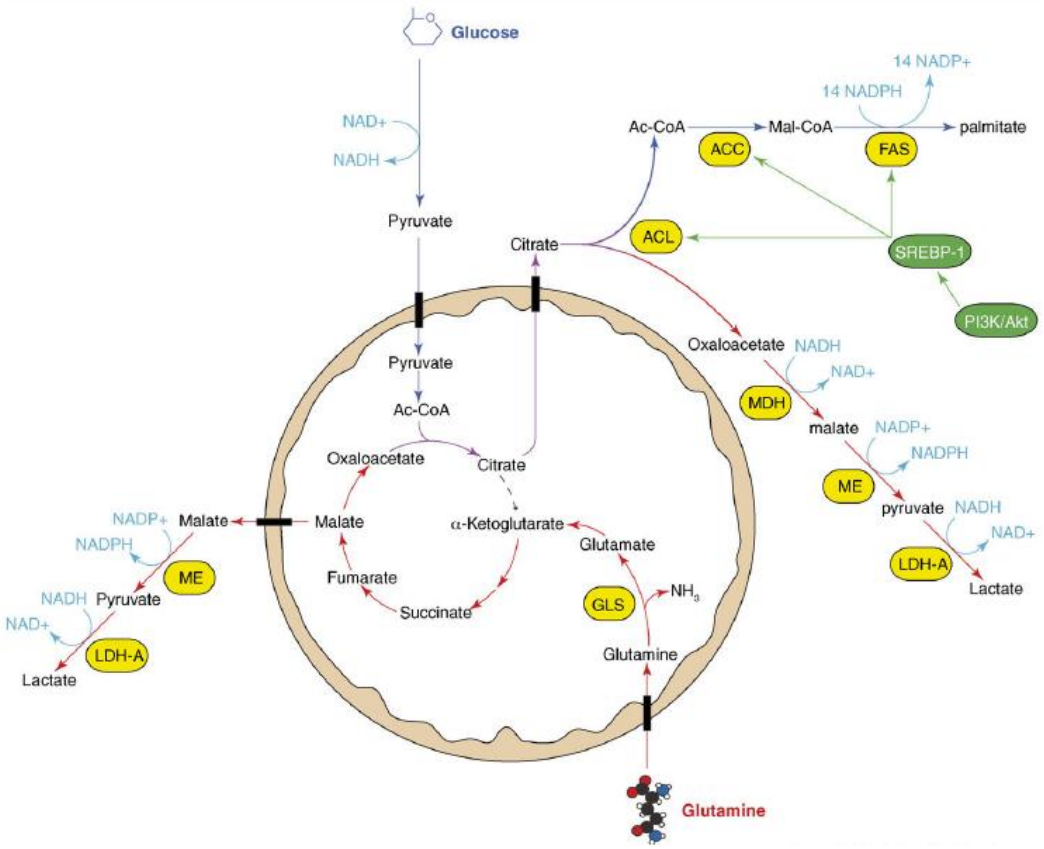


Figure 1.14 Glucose metabolism provides cells with a lipogenic precursor acetyl-CoA for fatty acid synthesis, which is enhanced in tumours due to oncogene-driven expression of ACL, ACC and FAS enzymes. (Reprinted from DeBerardinis et al. 2007, Copyright © 2008 with permission from Elsevier).

Consistent with high lipogenesis is the increased expression of lipogenic enzymes, observed in several tumour cells, namely ATP-citrate lyase (ACL, involved in the synthesis of acetyl-CoA from citrate), acetyl-CoA carboxylase (ACC, which catalyzes the production of malonyl-CoA from acetyl-CoA) and fatty acid synthase (FAS, a multi-enzyme complex catalysing fatty acid synthesis) (Swinnen et al. 2006). In fact, *in vitro* and/or *in vivo* inhibition of all these enzymes resulted in diminished cell proliferation, loss of cell viability or decrease of tumour (Kuhajda et al. 1994; Pizer et al. 1996; Hatzivassiliou et al. 2005; Brusselmans et al. 2005). It is established that tumour cells achieve rapid fatty acid synthesis through multiple effects of oncogenic mutations, particularly those involving the phosphatidylinositol-3 kinase PI3K/Akt/mTOR signalling pathway. This system stimulates expression of lipogenic genes through the activation of the sterol regulatory element binding protein-1 (SREBP-1), a transcription factor that targets ACL, ACC and FAS enzymes (Porstmann et al. 2005).

1.2.2 Exploiting tumour metabolism in cancer diagnosis and therapy

Knowledge of altered metabolism in cancer is at the basis of new diagnostic tools and anticancer drugs. One of the most emblematic examples, firmly implemented in the clinics, is ^{18}fluoro-2-deoxyglucose-positron emission tomography (FDG-PET) to image tumour development and regression. This technique, developed with basis on the 'Warburg effect', uses a radiolabelled glucose analogue to track its uptake, generally increased in several tumour types and shown to correlate with tumour aggressiveness (Gatenby and Gillies 2007). As a result, FDG-PET imaging has been successfully used in the diagnosis, staging and treatment monitoring of several cancers, including lung cancer (Zhu et al. 2011). In cases where tumours do not show pronounced 'Warburg effect', or when malignant lesions affect organs that normally take up (central nervous system) or excrete (kidneys, bladder and prostate) FDG, this method has, however, poor performance. Hence, other PET tracers have been developed for oncological purposes, focusing on metabolic pathways other than glycolysis. Examples include ^{11}C-acetate, a precursor of membrane fatty acids, and ^{18}F-choline, a substrate of choline kinase in choline metabolism, in prostate cancer (Grassi et al. 2012; Hausmann et al. 2014), ^{11}C-methionine, a precursor of adenosylmethionine required for polyamine synthesis, in brain tumour (Glaudemans et al.

2013), and ^{18}F-glutamine and ^{18}F-glutamate, targeting glutaminolysis (Lieberman et al. 2011; Ploessl et al. 2012).

Other tools under development for cancer diagnosis, based on altered tumour metabolism, include magnetic resonance spectroscopy imaging (MRSI) and dynamic nuclear polarization (DNP)-MRS. MRSI has expanded dramatically over the past decades as an adjunct technique to magnetic resonance imaging (MRI), being commonly applied to measure the levels of metabolites like *N*-acetylaspartate (NAA) in brain (Griffin and Kauppinen 2007), citrate in prostate (Swanson et al. 2006) and also choline compounds in breast (Bartella and Huang 2007), brain (Dowling et al. 2001) and prostate tumours (Seitz et al. 2009). DNP-MRS is a novel imaging technique that provides tens of thousands fold signal enhancement for stable isotope carbon-13 enriched compounds, making it possible to detect ^{13}C NMR signals from the substrate as well as from its metabolic products (Ardenkjaer-Larsen et al. 2003; Golman et al. 2003). Mainly due to its central role in cellular metabolism and favourable relaxation properties (Kurhanewicz et al. 2011), [1-^{13}C] pyruvate has been the most widely explored substrate in DNP-MRS. The first clinical trial of DNP-MRS using this substrate for the evaluation of prostate cancer is ongoing (Nelson et al. 2013), thus anticipating the rapid translation of this technique to clinical research and even clinical practice (Kurhanewicz et al. 2011).

Research on altered tumour metabolism has also allowed novel therapeutic targets to be identified, stimulating the development of multiple anticancer drugs targeting metabolic enzymes with key roles in tumour growth. A list comprising examples of such potential therapeutic compounds, together with their main targets, is shown in Table 1.2.

Table 1.2 Examples of anticancer strategies targeting glucose, glutamine and purine metabolisms. (Adapted from Madhok et al. 2011). 2DG: 2-deoxyglucose; 3PO: 3-(3-pyridinyl)-1-(4-pyridinyl)-2-propen-1-one; L-DON: 6-diazo-5-oxo-L-norleucine; 5-FU: 5-fluorouracil; NCT: National Clinical Trial.

Strategy	Compound	Drug development	Reference/NCT number
Inhibition of glycolytic enzymes			
Hexokinase	Lonidamine	Clinical trials, approved in Europe	Gatzemeier et al. 1991; Gadducci et al. 1994; De Lena et al. 2001
	2DG	Clinical trials	Mohanti et al. 1996; Raez et al. 2013
6-Phoshofructo-1-kinase	3PO	Preclinical	Clem et al. 2008
Pyruvate kinase	TLN-32	Clinical trials	Thallion Pharmaceuticals 2009

Table 1.2 (continued)

Strategy	Compound	Drug development	Reference/NCT number
Inhibition of pentose phosphate pathway			
Transketolase	Oxythiamine	Preclinical	Raïs et al. 1999
Promotion of oxidative phosphorylation			
Inhibition of lactate dehydrogenase	RNA interference	Preclinical	Xie et al. 2009
Attenuation of HIF1 activity			
Inhibition of HIF1α translation	Topotecan	Approved by FDA, clinical trials	NCT00770536, NCT00698516, NCT00770731
Inhibition of HIF1α stability	YC-1	Preclinical	Yeo et al. 2003; Kim et al. 2006
Inhibition of glutamine metabolism			
Glutaminase	L-DON	Clinical trials	Ahluwalia et al. 1990; Mueller et al. 2008
Inhibition of purine synthesis			
Thymidylate synthase	5-FU	Clinical trial	Garcia et al. 2003; O'Connell et al. 2006
Dihydrofolate reductase	Methotrexate	Registered	

Targeting glucose metabolism is one of the most explored strategies, comprising different approaches: i) inhibition of glycolytic enzymes, generally aimed at depriving tumour cells of ATP; ii) inhibition of the PPP, with a view to decrease the production of NADPH (which plays an important role in preventing oxidative stress) and of pentoses (required for nucleic acids synthesis); iii) promotion of oxidative phosphorylation, in order to shift the cells from the dependence on glycolysis and reactivate apoptosis through the mitochondrial pathway; and iv) attenuation of HIF1 activity, a critical upstream regulator of glucose metabolism. Other cancer therapy targets extensively studied are the glutamine metabolism and the purine synthetic pathways, as exemplified in Table 1.2.

1.3 Metabolomics in clinical oncology

Before the last decade, knowledge of altered tumour metabolism was largely inferred from gene and protein expression changes. However, this approach does not account for factors such as post-translational modifications, allosteric regulation or metabolic compartmentalization, eventually providing an incomplete picture of the malignant metabolic phenotype. The technological developments employing advanced analytical techniques, such as nuclear magnetic resonance (NMR) spectroscopy or mass spectrometry

(MS), have enabled the direct assessment of the metabolic composition of biological tissues and fluids, thereby providing unique and complementary insights into tumour metabolism. The so called metabolomics approach has thus entered the 'omics' world to nicely complement genomics, transcriptomics and proteomics (Figure 1.15), by offering a direct window into potential metabolic diagnostic markers or therapeutic targets, and eventually allowing a more comprehensive and interpretable picture of cellular phenotypes to be obtained. The general strategy employed in metabolomics, as well as the fundamentals of the main analytical and statistical tools used will be described in the next sections.

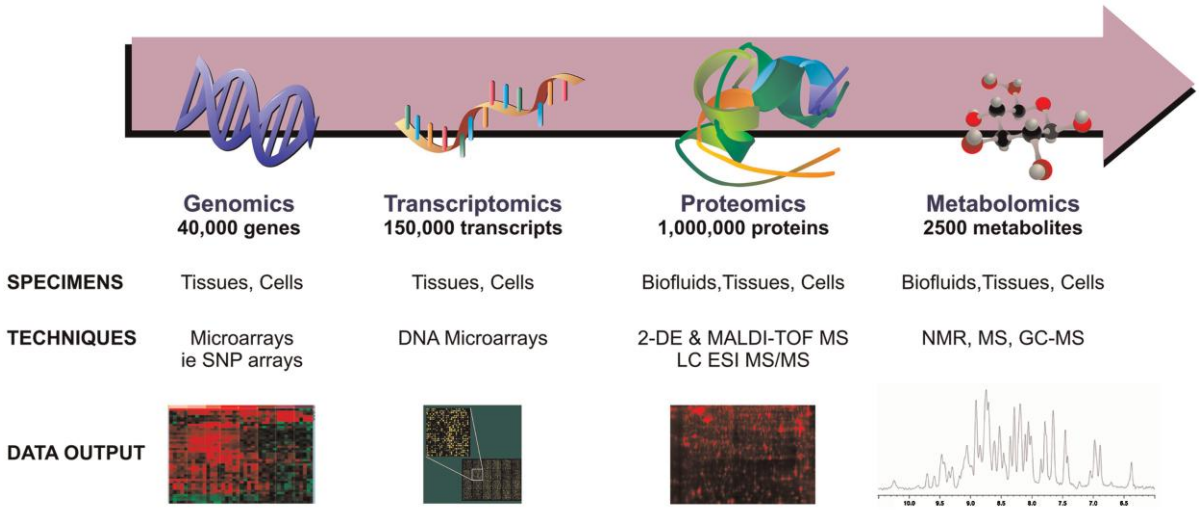


Figure 1.15 Comparison between omic sciences. From Davis et al. 2011, Copyright © 2010 Wiley-Liss, Inc.

1.3.1 The metabolomics approach: concept and strategy

Metabolomics has initially been described as 'the comprehensive identification and quantification of all metabolites in a biological system' (Fiehn 2002), while the similar term metabonomics, devised by Nicholson et al. in the late 1990s, has been defined as 'the quantitative measurement of the dynamic multiparametric response of living systems to pathophysiological stimuli or genetic modification' (Nicholson et al. 1999). Although, historically, there are some conceptual differences between the two terms, today they are often used interchangeably to express the fluctuations on the inventory of endogenous metabolites (the metabolome) upon a given perturbation (e.g., disease, environment,

dietary or therapeutic regimens, genetic modification). The term metabolomics will be used hereafter in this thesis.

Metabolites are small molecules (typically with less than 1 kDa) involved in intermediary metabolism, their levels depending on (and affecting) upstream modifications in genes and protein expression. Metabolite abundances are also modulated by several factors unrelated to genome, such as interaction with commensal microorganisms, nutritional and other lifestyle-related aspects. Therefore, by reflecting the complex interplay between the genome (and encoded proteins) and the environment, the metabolome closely expresses the overall physiological status of an organism. By having the ability to capture a comprehensive, non-selective picture of variations in the metabolome, metabolomics has great potential for revealing new insights into disease biochemistry or identifying possible disease markers.

Different types of samples may be used in metabolomic studies, including cultured cells, tissues and several biofluids (e.g., urine, blood plasma/serum, bile, cerebrospinal fluid). Generally, the strategy followed, depicted in Figure 1.16, involves: i) sample collection, ii) sample analysis using one or more analytical platforms (mainly ^{1}H NMR spectroscopy and hyphenated MS), iii) assignment of spectral data, and iv) application of bioinformatic tools to maximise information recovery from the complex datasets produced.

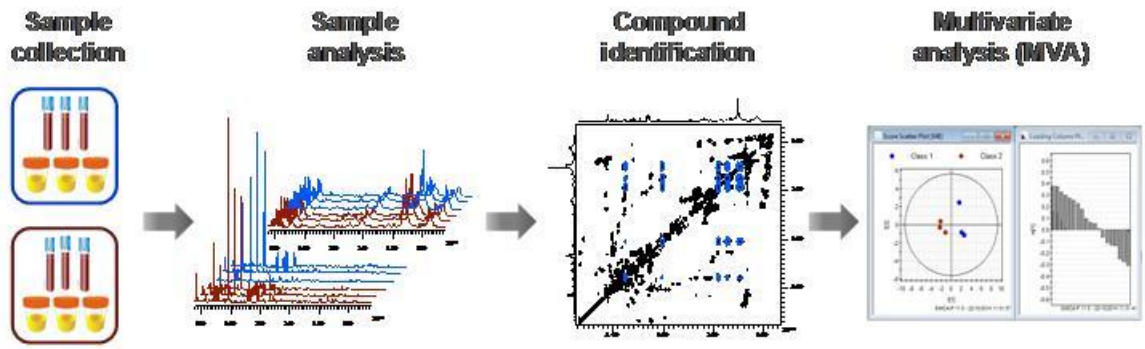


Figure 1.16 Typical workflow diagram applied in metabolomic studies.

The choice of the analytical platform greatly depends on the analysis' purposes, the type of compounds targeted and the biological specimen to be analysed. In spite of being inherently less sensitive than MS, NMR allows the direct analysis of tissue/cells (by HRMAS, described in section 1.3.2.4), it is highly reproducible, non-destructive (allowing for sample recovery), and gives non-selective, qualitative and quantitative information on

the main metabolites present (with a limit of detection typically in the sub-milimolar range, Wishart 2013). On the other hand, MS-based methods like LC-MS or GC-MS generally have greater sensitivity, with detection limits reaching <1 nM (Wishart 2013), thus enabling the detection of hundreds to thousands of compounds. However, the number of molecules actually identified is usually much inferior, as automated approaches of chemical identification are still lacking. Moreover, although analyte quantification is possible, it can be impaired by variable ionization and ion suppression effects (Lindon et al. 2007). Finally, MS studies must carefully consider reproducibility, as it may be affected by changes in response over time, arising from the physical interaction of samples with the instrument (Dunn et al. 2011).

In this project, ^{1}H NMR spectroscopy was the main platform used for the analysis of tissue, plasma and urine samples, while MS (coupled to liquid chromatography) was only preliminarily explored in urine analysis. The following sections describe the basic principles of these techniques and how they can be employed in metabolomic studies.

1.3.2 Nuclear Magnetic Resonance (NMR) spectroscopy

1.3.2.1 Basic principles

Quantum physics postulates that the nuclei of certain atoms possess an angular momentum, P, which in turn is responsible for originating a small magnetic dipole moment, μ. The relationship between magnetic moment and angular momentum is given by Equation 1.1.

$$\mu = \gamma P \qquad \textbf{Equation 1.1}$$

where γ (in rad·T^{-1}·s^{-1}) is the magnetogyric ratio, an isotope-specific constant. Angular momentum and nuclear magnetic moment are quantized (can only adopt discrete values), so, in the *z*-direction of an arbitrarily chosen Cartesian coordinate system, they can be defined as follows (Equation 1.2 and Equation 1.3),

$$P_z = \hbar m_z \qquad \textbf{Equation 1.2}$$

$$\mu_z = \gamma \hbar m_z \qquad \textbf{Equation 1.3}$$

with m_z being the magnetic quantum number and $\hbar=h/2\pi$ (where h is Planck's constant). The magnetic quantum number is, in turn, defined as m_z=-I, -I+1, -I+2, ..., I, where I is the nuclear spin quantum number. I may be equal to zero (nuclei with even number of protons and neutrons, e.g., ^{12}C, not detectable by NMR) or multiple of ½ (either integer, in the case of isotopes with odd number of protons and neutrons, or half-integer values). For a spin of quantum number $I\neq0$, there exist 2I+1 possible spin orientations (nuclear spin states) relative to an axis. In the case of a nucleus with spin ½ (^{1}H, ^{13}C, ^{15}N, ^{19}F and ^{31}P), there are two possible orientations: the spin state with m_I=+½, denoted α, and the spin state with m_I=-½, denoted β. Both α and β states have the same energy, that is, they are degenerate. However, when placed in an external static magnetic field, B_0, this degenerancy is lifted as a result of the interaction of the nuclear magnetic moment μ with B_0 (Figure 1.17).

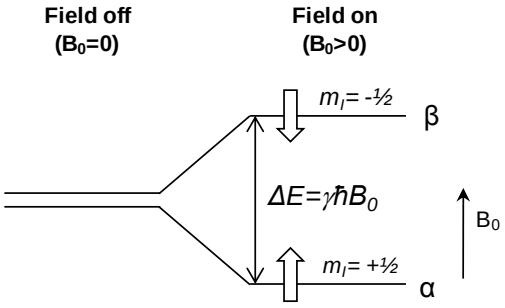


Figure 1.17 Schematic representation of the nuclear spin energy levels of a spin-½ nucleus in a magnetic field.

The energy difference between the two states (E_β-E_α), given by Equation 1.4, is proportional to the strength of B_0 and is called the Zeeman effect.

$$\Delta E = \gamma \hbar B_0 \qquad \textbf{Equation 1.4}$$

The α and β states are filled according to the Boltzmann distribution (Equation 1.5), where $N_{\alpha,\beta}$ represents the number of nuclei in each spin state, ΔE is the energy difference between the two states, k is the Boltzmann constant (1.381×10^{-23} J·K^{-1}) and T is the absolute temperature.

$$\frac{N_\beta}{N_\alpha} = e^{-\Delta E/kT} = e^{-\gamma h B_0/2\pi kT} \approx 1 - \frac{\gamma h B_0}{2\pi kT} \qquad \textbf{Equation 1.5}$$

Since ΔE is very small, at equilibrium, there will be a slight excess of nuclei in the lower energy α state, compared to the β state, giving rise to a macroscopic magnetic moment, the so-called magnetization, M_0 ($M_0 = \sum_i \mu_i$). This small energy difference between spin states explains the inherently low sensitivity of NMR, which can be increased by raising the magnetic field strength or by lowering the temperature (Equation 1.5) (Günther 2013).

The effect of a static magnetic field can also be described as the magnetic moment of an individual spin precessing about the axis of the external magnetic field B_0 (Larmor precession) (Figure 1.18a), with a certain angular velocity, ω_0, called the Larmor frequency (Equation 1.6), which divided by 2π, gives the resonance frequency in Hz, ν_0 (Equation 1.7). ν_0 is thus the basic frequency of a specific isotope at a given field strength.

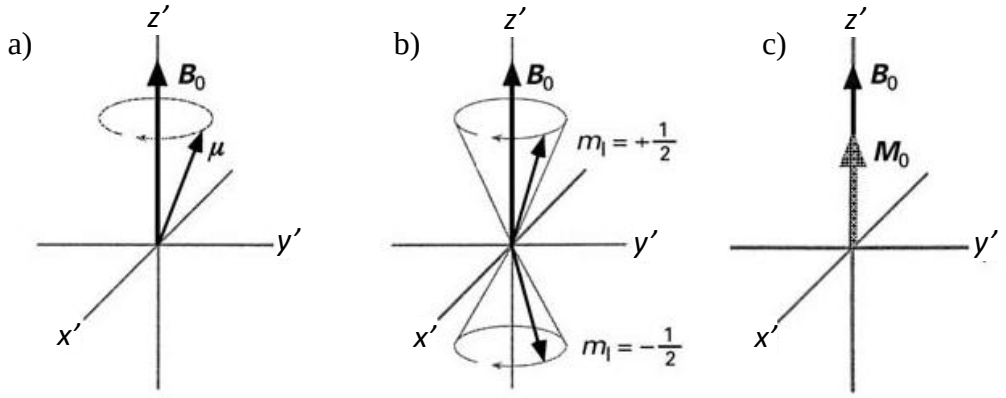


Figure 1.18 a) Precession of an individual magnetic moment μ about the external magnetic field B_0. b) Precession of magnetic moments in the α (m_I=+½) and β (m_I=-½) spin states. c) Resultant magnetic moment M_0 of a large number of equivalent nuclei, showing a small excess population of nuclei in the α spin state. (Adapted from Reynolds 1999).

$$\omega_0 = -\gamma B_0 \qquad (\text{rad}\cdot\text{s}^{-1}) \qquad \textbf{Equation 1.6}$$

$$\nu_0 = \frac{|\gamma| B_0}{2\pi} \qquad (\text{Hz}) \qquad \textbf{Equation 1.7}$$

In simple terms, to observe an NMR signal, the energy necessary to stimulate the transition between states must match the resonance frequency, thus meeting the resonance condition. In an NMR experiment, that energy is provided by applying radiofrequency (RF) radiation for a short period (RF pulse, typically of the order of μs) in the transverse plane (perpendicular to B_0) (Figure 1.19). This pulse is produced by a small coil

surrounding the sample, its magnetic component B_1 oscillating at the Larmor frequency of the spins. As the RF pulse is applied, the net magnetization M_0 is tilted away from alignment with B_0 toward the *x'y'*-plane (a rotating frame is used to simplify visualization of concomitant rotating fields and precessing vectors). Once the RF pulse is off, the magnetization in the transverse plane gradually disappears and is re-established along the *z*-axis, through a relaxation process (further described below). Precession of the magnetization in the *x'y'*-plane induces an oscillating signal in the detection coil, designated free induction decay (FID). This time-domain signal is then transformed into the frequency-domain spectrum through Fourier transform (FT).

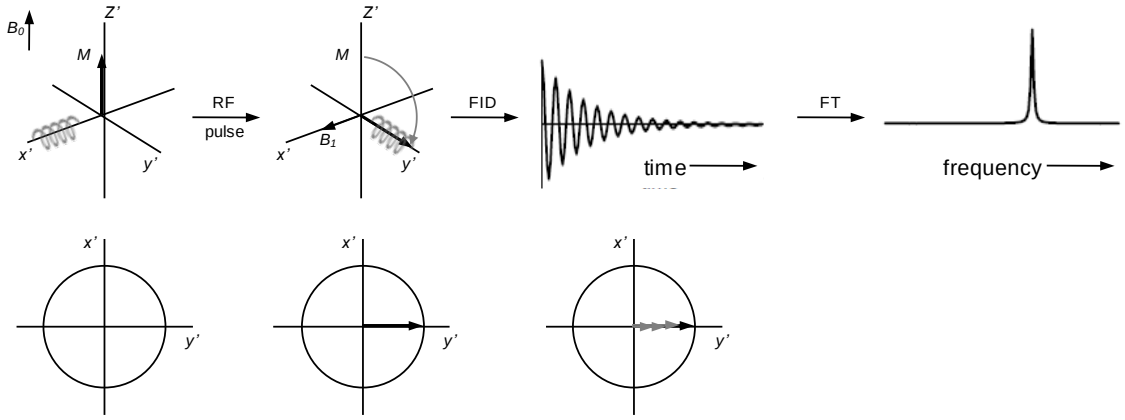


Figure 1.19 Schematic representation of the basic steps of an NMR experiment. (Adapted from Keeler 2010).

1.3.2.2 Chemical shift and scalar coupling

The electrons surrounding the nucleus create a local magnetic field (B_{loc}), opposite to B_0, which causes a shielding effect and makes the effective magnetic field (B_{eff}) experienced by the nucleus to be:

$$B_{eff} = B_0 + B_{loc} = B_0(1 \pm \sigma) \qquad \textbf{Equation 1.8}$$

where σ is a shielding constant for a given nucleus at a particular position in the molecule (Jacobsen 2007). Thus, nucleus with different chemical environments (non-equivalent) will feel different effective fields and will resonate at slightly different frequencies, giving rise to different positions in the NMR spectrum. These positions are indicated by chemical shift

values (δ), in parts per million (ppm), expressed by the resonance frequency of a given nucleus (ν), relative to a standard (e.g., tetramethylsilane, TMS) and normalized with respect to the frequency of that standard (ν_{ref}) (Equation 1.9).

$$\delta = \frac{\nu_x - \nu_{ref}}{\nu_{ref}} \times 10^6 \qquad \textbf{Equation 1.9}$$

In this way, chemical shifts obtained from different instruments can be compared, being characteristic for a certain nucleus in a specific compound, in a particular solvent, regardless of the operating frequency of the spectrometer. Generally, more shielded nuclei resonate at lower frequencies (lower chemical shifts), whereas less shielded nuclei, such as those with electronegative neighbours (e.g., N, O, Cl), resonate at higher frequencies (higher chemical shift). This valuable property has been used since the early days of NMR to retrieve structural information on molecules. Figure 1.20 shows the typical ^{1}H chemical shift ranges of some functional groups in organic compounds.

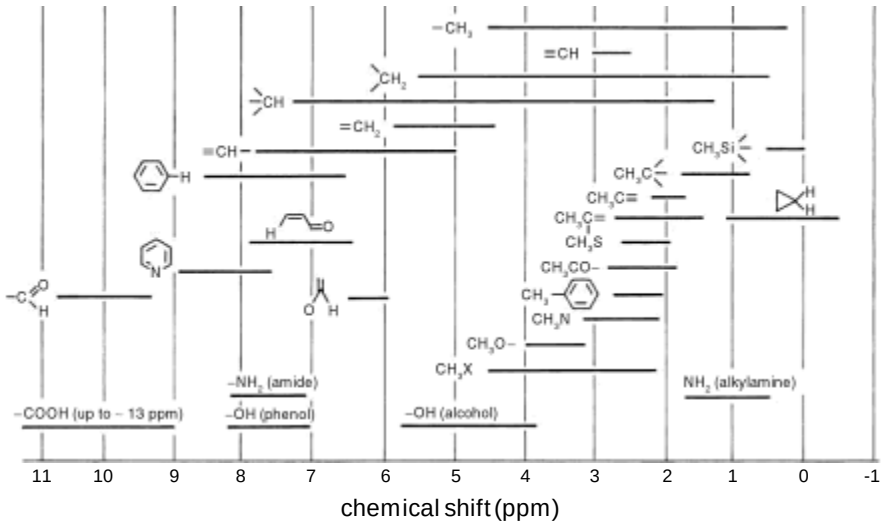


Figure 1.20 Typical ^{1}H chemical shift ranges (in ppm) of some functional groups. (Adapted from Günther 2013).

Another important feature of NMR signals is their splitting pattern, which reflects the interaction between neighbouring spins, via the electrons in the chemical bond – a phenomenon designated scalar, spin-spin or *J*-coupling. For nuclei with spin quantum number I=½, the multiplicity of the line splitting equals n+1, where n is the number of non-

equivalent nuclei in the neighbouring group, and the line separation (in Hertz) corresponds to the coupling constant, *J*, between the nuclei under consideration. The magnitude of *J* generally decreases as the number of bonds between the coupled nuclei increases. Coupling between like spins (same isotope) is called homonuclear coupling and that between different types of isotopes is called heteronuclear coupling.

1.3.2.3 Spin relaxation

There are two different processes by which nuclear spins return to the thermal equilibrium: longitudinal (or spin-lattice) relaxation and transverse (or spin-spin) relaxation. Longitudinal relaxation is responsible for the recovery of magnetization along the *z*-axis, the energy lost by the spins being transferred into the surroundings in the form of heat. The longitudinal relaxation efficiency is characterised by a time constant T_1 and its behaviour as function of time is described by Equation 1.10.

$$M_z(t) = M_0\left(1 - e^{-\frac{t}{T_1}}\right) \qquad \textbf{Equation 1.10}$$

Transverse relaxation, on the other hand, describes the fanning-out (loss of phase coherence) of spins in the *x'y'*-plane as they precess at different rates, leading to no net magnetization in the transverse plane. Dipolar interactions with nearby spins are the main contributor for this relaxation mechanism, hence named spin-spin relaxation. Again, this type of relaxation is assumed to occur with an exponential decay, characterised by the transverse relaxation time constant T_2 (Equation 1.11).

$$M_{xy}(t) = M_0 e^{-t/T_2} \qquad \textbf{Equation 1.11}$$

In an NMR experiment, T_1 values should be taken into account to define the interscan delay (recycle time), as incomplete recovery of *z*-magnetization between scans will lead to signal saturation (non-quantitative conditions). In order to have quantitative conditions the recycle time should be at least five times the longest T_1 (99% of magnetization recovered, Bharti and Roy 2012). T_2 values, in turn, determine how fast the signal decays and thereby influence the signal line width. As the line width is proportional

to $1/\pi \cdot T_2$, molecules with shorter T_2 display broader lines. Another important contribution to the decay of the signal and, hence, to line broadening, is the B_0 inhomogeneity, taken into account by specifying the effective transverse relaxation time constant T_2^* (Equation 1.12), where $T_{2homog.}$ refers to T_2 relaxation in a perfectly homogenous B_0 field.

$$\frac{1}{T_2^*} = \frac{1}{T_{2homog.}} + \frac{1}{T_{2inhomog.}} \qquad \textbf{Equation 1.12}$$

The different dependencies of T_1 and T_2 with respect to overall molecular tumbling rate or correlation time (τ_c) is shown in (Figure 1.21). For small and medium-sized organic molecules in solution (with rapid tumbling, hence short τ_c), ^{1}H T_1 values are typically of the order of seconds and T_2 values are similar to corresponding T_1's. For macromolecules (with larger τ_c), T_2 values are generally much smaller and can be of a few milliseconds. Increasing solvent viscosity or reducing sample temperature (hence, decreasing tumbling rates) can also reduce T_2 time constants, thus broadening NMR resonances.

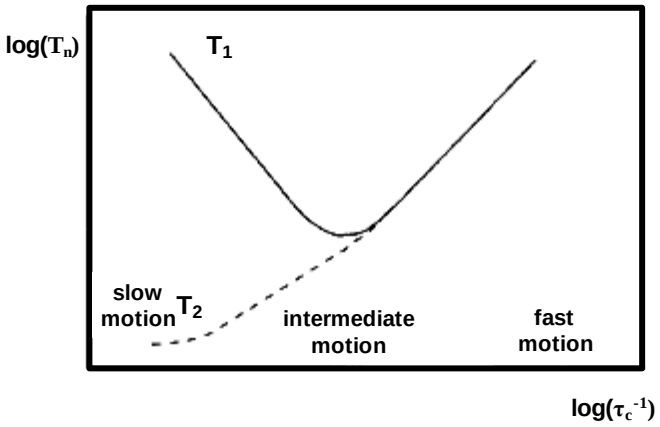


Figure 1.21 Dependence of T_1 and T_2 relaxation time constants as function of correlation time, τ_c (the average time it takes a molecule to rotate one radian). (Adapted from Claridge 2000).

1.3.2.4 High-resolution magic angle spinning (HRMAS)

In the solution state, where molecules have fast isotropic motions, nuclear interactions are dominated by the Zeeman effect (interaction with the external static magnetic field, described above), and chemical shifts and coupling constants show discrete average values. Conversely, in the solid state, where molecular motion is restricted due to

the relatively fixed orientation of molecules, additional interactions, which contribute to broaden spectral lines, have to be considered. These include, chemical shift anisotropy (CSA, resulting from non-spherical electron distribution), through space dipole-dipole interactions and, for nuclear spins with I >½, quadrupolar interactions. The typical magnitudes are shown in Table 1.3. Dipolar couplings are on the order of many tens of kHz and CSA can also cover hundreds of parts per million, thus leading to very broad lines.

Table 1.3 Typical magnitude of nuclear spin interactions in the solid state (Fyfe 1983).

Interaction	**Typical magnitude**
Zeeman	1-1000 MHz
Quadrupolar	0-1000 MHz
Chemical shift	0-100 kHz
Dipole-dipole	0-100 kHz
Scalar coupling	0-10 kHz

In the late 1950s, a technique was devised to minimise the above mentioned anisotropic interactions and allow NMR signals of solids to be recovered with improved resolution (Andrew and Newing 1958; Lowe 1959). This technique, called magic angle spinning (MAS), consists of rapidly spinning the sample at an angle of 54.74º (the magic angle) relative to the applied magnetic field (Figure 1.22a). For ½-spin nuclei, the main anisotropic interactions (CSA and dipolar couplings) have an angular dependence of $(3\cos^2\theta\text{-}1)/2$, where θ is the angle between the static magnetic field (B_0) and the internuclear vector responsible for the interaction (Figure 1.22b).

For example, the Hamiltonian operator for the dipolar coupling may be written as:

$$\mathcal{H}_D^{ij} = \frac{h\,\gamma_i\gamma_{j\,ij}}{4\pi^2 r_{ij}^3}\frac{3cos^2(\theta_{ij}-1)}{2}(I_iI_j - 3I_{iz}I_{jz}) \qquad \textbf{Equation 1.13}$$

where γ_j and γ_j are, respectively, the magnetogyric ratio constants of spins *i* and *j*, *h* is Planck's constant, r_{ij} is the distance between the two coupled nuclei, θ_{ij} is the angle the connecting vector makes to the external field, and *I* is the spin-operator.

Spinning the sample at an angle β relative to the magnetic field imposes an additional dependence on the factor $(3\cos^2\beta\text{-}1)/2$, which is zeroed when β=54.74º, causing anisotropic interactions to be averaged out (as long as spinning rates are high enough) and spectral lines to be significantly narrowed.

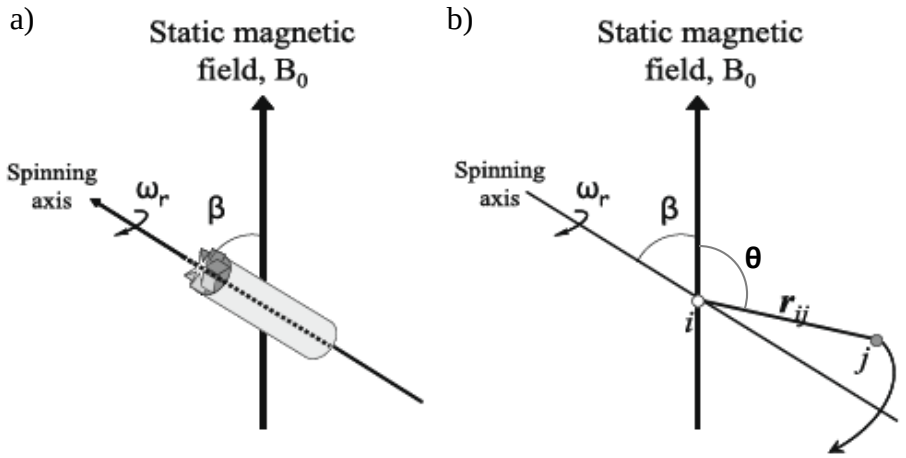


Figure 1.22 Schematic representation of a) MAS rotor in a magnetic field, B_0, where β is the magic angle 54.7º, and ω_r is the spinning rate. b) Two nuclei *i* and *j* in a solid under the same conditions as in a), where r_{ij} is the internuclear distance. (Adapted from Sitter et al. 2009).

For semi-solid samples, such as tissue biopsies, which retain a great deal of molecular motion, the anisotropic interactions referred above are partially averaged and their effect on line broadening greatly reduced. For ^{1}H nuclei, chemical shift anisotropies are generally rather small, hence the major causes for line broadening in the ^{1}H NMR spectra of tissue samples are residual dipolar couplings (of a few hundred Hertz) and anisotropic susceptibility effects. These effects consist of variations in local magnetic susceptibility due to the heterogeneous nature of the sample, as different sample regions will be magnetized to different extents by the applied magnetic field. This magnetic susceptibility inhomogeneity (which in proton spectra is usually of the order of a few tens of Hz) has also an angular dependence on the factor $3\cos^2\theta\text{-}1$, and, thus, can be greatly attenuated by MAS. Normally, removal of line broadening effects in tissues using MAS requires a spinning rate of a few hundred Hz, but this would produce spinning side bands (SSB) near the central region of the spectrum, which could obscure metabolite signals of

interest. Therefore, spinning rates that allow SSB to be out of the region of interest are typically of a few kHz (e.g., 4 kHz for 400 MHz ^{1}H observation).

The analysis of intact tissues (and other biological samples) has also significantly benefited from the development of high resolution probes allowing magic angle spinning (HRMAS probes). This type of probes enables a lock system to ensure long-term stability of the magnetic field and pulsed field gradients to be used, allowing for pulse sequences from liquid state NMR to be applied without modifications. In this thesis, ^{1}H HRMAS was used to record well-resolved spectra from intact lung tissues, as will be shown in Chapter 3.

1.3.2.5 Main one- and two-dimensional NMR experiments

Acquisition of one-dimensional (1D) ^{1}H NMR spectra of biological samples typically requires the use of solvent suppression schemes, as the dominant water signal would obscure the unsuppressed spectrum and cause dynamic range problems. One of the most popular experiments is the 1D NOESY-presat sequence (named as 'standard 1D' throughout this thesis), as it provides good water suppression while maintaining a flat baseline. In this pulse sequence, the water resonance is saturated by selectively irradiating it during the interscan relaxation delay (RD) and the mixing time (τ_m) of the pulse sequence (Figure 1.23).

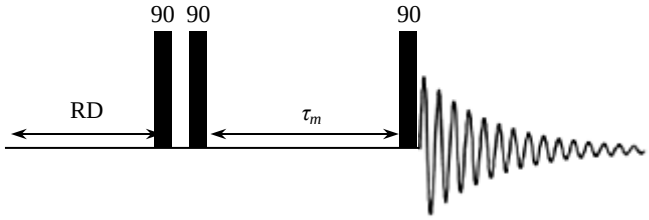


Figure 1.23 Schematic illustration of the 1D NOESY-presat pulse sequence. The solvent is presaturated by application of a weak RF field during the interscan relaxation delay (RD) and mixing time (τ_m). (Adapted from Zerbe and Jurt 2014).

For samples with large superimposition of signals from small metabolites and larger molecules (e.g., tissues and blood plasma), additional 1D experiments are useful, in order to selectively observe those two types of compounds. The T_2-edited Carr-Purcell-Meiboom-Gill (CPMG, Meiboom and Gill 1958) experiment (schematically represented in

Figure 1.24) is usually recorded to attenuate the signals of fast-relaxing macromolecules, improving the visibility of narrow signals from small metabolites. This sequence uses a 90º pulse followed by a spin-echo period (delay-180º-delay), repeated *n* times. During the total echo time ($2\tau n$), M_{xy} decays at a rate that is inversely proportional to T_2. Thus, by choosing an appropriate echo time, the signals of macromolecules (with shorter T_2's) can be attenuated in relation to those of small molecules with slower T_2-relaxation.

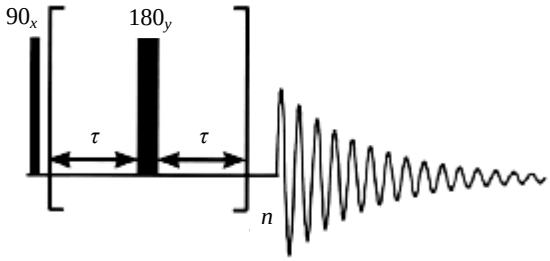


Figure 1.24 Schematic illustration of the Carr-Purcell-Meiboom-Gill (CPMG) pulse sequence. (Adapted from Zerbe and Jurt 2014).

Oppositely, the 1D diffusion-edited experiment is used to retain the signals of large macromolecules (e.g., lipids, proteins) or compounds with restricted mobility. The principle behind diffusion-editing lies in the application of a first gradient to dephase magnetization, followed by a delay time during which molecules diffuse, and a later gradient (of opposite sign to the initial) that refocuses magnetization. Those molecules which diffuse faster (smaller molecules) will have lost magnetization after the delay time, whereas slow-diffusing molecules will preserve their signals in the spectrum.

One-dimensional ^{1}H spectra of biological samples are typically characterised by hundreds of overlapping peaks, their unambiguous assignment requiring the use of two-dimensional (2D) experiments, as 2D spectra show increased signal dispersion and are able to provide information on connectivities between signals. 2D NMR experiments are composed of four building blocks: preparation (or excitation), evolution, mixing and detection (Figure 1.25a). In the preparation time, the sample is excited by one or more pulses, depending on the sequence, and the resulting magnetization is allowed to evolve during the time period t_1. Then, further pulse(s) is(are) applied during the mixing time (transfer of magnetization from one spin to another, via scalar or dipolar coupling, or via exchange processes), after which the signal is recorded as a function of the second time

variable, t_2 (analogous to the detection period of any 1D experiment). The entire experiment is repeated with step-wise incrementation of t_1, to allow monitoring the evolution of the prepared spin state. Fourier transformation with respect to t_2 will result in a set of spectra in which signal intensities are modulated as a function of t_1. FT with respect to t_1 converts the frequency modulation into peaks in the 2D spectrum (Figure 1.25b).

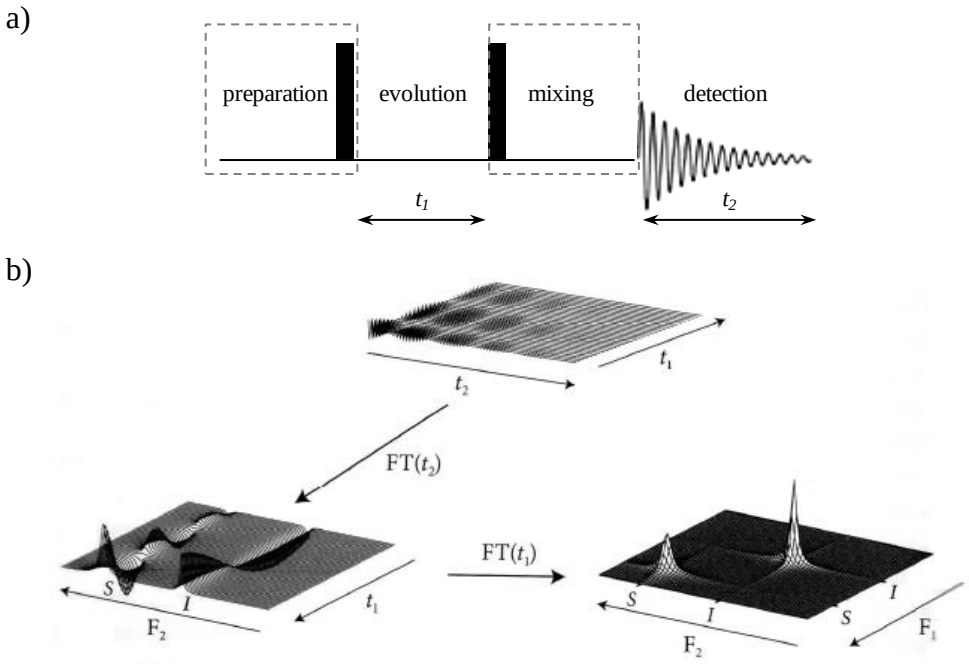


Figure 1.25 a) Basic building blocks of 2D NMR experiments. b) Schematic illustration of the principle behind 2D NMR spectroscopy. (Adapted from Zerbe and Jurt 2014).

Some of the most commonly used 2D experiments in metabolomics are those revealing homonuclear connectivities, such as ^{1}H-^{1}H correlation spectroscopy (COSY) and total correlation spectroscopy (TOCSY), which rely on scalar couplings between protons. In particular, COSY detects couplings between protons that are 2-3 bonds apart, whereas TOCSY further shows connectivities for protons that are up to 5 or more bonds apart, allowing for complete spin systems to be assigned (Jacobsen 2007). Detection of long range correlations in the TOCSY experiment relies on the coherence transfer between spins at a temporary state of strong coupling, which is achieved by using a mixing (spin-lock) sequence. A typical spin-lock sequence is the Malcolm Levitt (MLEV) sequence, which comprises 16 composite pulses followed by a regular uncompensated 180º pulse (Bax and Davis 1985). A schematic representation of a TOCSY experiment incorporating in the

mixing period an MLEV-17 spin-lock and two trim pulses (at the beginning and end of the mixing period to defocus any magnetization that is not parallel to the *x* axis) is shown in Figure 1.26.

Figure 1.26 Schematic illustration of the TOCSY experiment based on the MLEV-17 mixing scheme. SL: spin-lock trim pulses. (Adapted from Claridge 2000).

Heteronuclear correlation experiments, namely those showing ^{1}H-^{13}C correlations, can also be very useful for assignment purposes. One of the most frequently used is the heteronuclear single-quantum coherence (HSQC) experiment, whereby the chemical shift of ^{13}C is correlated to the chemical shift of the directly bound proton (via ^{1}H-^{13}C scalar coupling) (Bodenhausen and Ruben 1980). This experiment is based on the transfer of magnetization from the proton to the heteronucleus using an INEPT (insensitive nuclei enhancement by polarization transfer) block, followed by the recovery of magnetization from the carbon to the proton before acquisition (Figure 1.27). This allows the detection of proton magnetization, instead of carbon's, making HSQC more sensitivity than other heterocorrelation pulse sequences (like HETCOR).

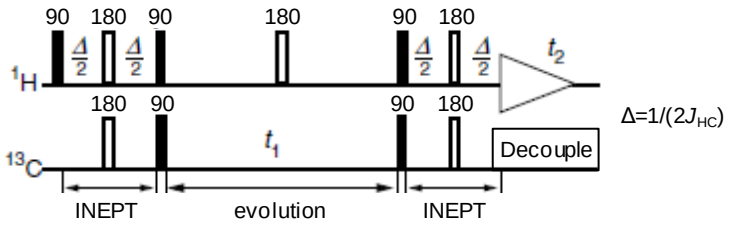


Figure 1.27 Schematic representation of the ^{1}H-^{13}C HSQC pulse sequence. t_1 is the evolution time and t_2 is the acquisition time, and J_{HC} is ca. 150 Hz. (Adapted from Keeler 2010).

Another useful 2D experiment is the *J*-resolved (JRES), which separates chemical shifts and scalar couplings into F1 and F2 dimensions, respectively (Ludwig and Viant 2010). This enables spectral complexity to be significantly reduced and information on signal multiplicity and coupling constants to be more easily extracted. One of the most

basic pulse sequences of 2D JRES is illustrated in Figure 1.28a. After the 90º pulse to create transverse magnetization, the chemical shift evolution is refocused by applying a 180º pulse, while scalar coupling remains active, so that the F1 dimension only contains coupling information. On the other hand, during t_2, both chemical shift and *J*-coupling evolution are active, resulting in proton multiplets in the F2 dimension (Figure 1.28b), which complicate spectral interpretation. A processing procedure consisting of tilting the multiplets by 45º about their midpoints is then applied, to retain only the chemical shift information in the F2 dimension (Figure 1.28c).

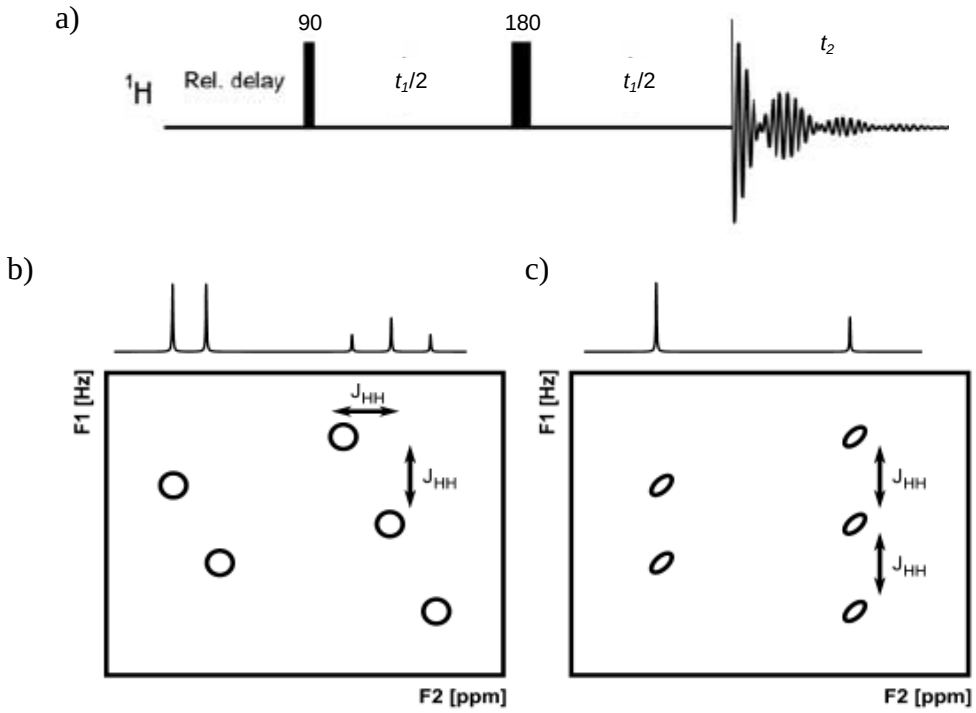


Figure 1.28 a) Schematic representation of the 2D ^{1}H J-resolved pulse sequence, where t_1 is an incremented time delay and t_2 is the acquisition time. b) and c) Schematic representation of a 2D JRES NMR spectrum after Fourier transform (b), and after tilting the spectrum by 45º (c), where J_{HH} is the *J*-coupling constant between protons. (Adapted from Ludwig and Viant 2010).

1.3.3 Mass Spectrometry (MS)

1.3.3.1 Basic principles

The core principle of mass spectrometry is the measurement of mass-to-charge ratios (*m/z*) of ionised molecules, clusters of molecules, complexes or fragments. Simplistically, the mass spectrometer comprises the following elements: an ion source, a mass analyser, a detector and a data system (Figure 1.29). In general, the analytes are ionised (either at high

vacuum pressure or at atmospheric pressure) and transferred (in the gas phase) to the high vacuum region of the mass spectrometer, where they are separated according to their *m/z* and detected. The separation and detection steps take place under high vacuum pressure to reduce the number of ion-ion and ion-molecule collisions, which can influence the mass resolution, mass accuracy and sensitivity of instruments (Dunn et al. 2011). The resulting mass spectrum is a plot of *m/z* vs. signal intensity (reflecting ion abundance) (Hoffmann and Stroobant 2007).

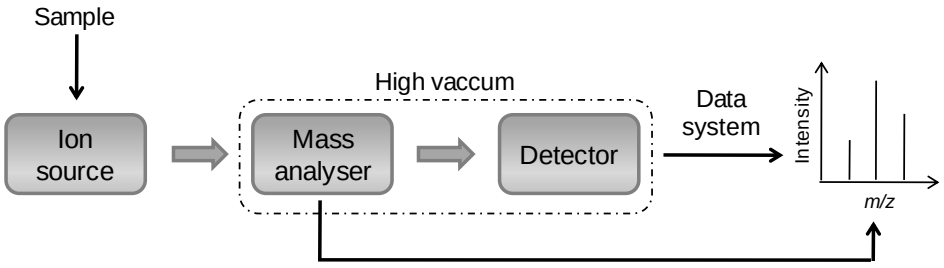


Figure 1.29 Basic components of an Electrospray Ionisation (ESI) mass spectrometer. (Adapted from Wang and Griffiths 2008).

A number of different types of ionisation methods can be employed for generating gas-phase ions, including: electron ionisation (EI), chemical ionisation (CI), electrospray ionisation (ESI), atmospheric pressure chemical ionisation (APCI) and desorption electrospray ionisation (DESI) (Wang and Griffiths 2008). Also, mass analysers can be of different types, the most popular in metabolic profiling studies being the linear quadrupole, the quadrupole ion-trap and the time-of-flight (ToF) analysers (Wang and Griffiths 2008). In the present work, the analysis of urine samples by hyphenated mass spectrometry was performed by using electrospray ionisation (ESI) coupled to a ToF analyser; therefore these methods will be described in more detail.

Electrospray is a commonly used ionisation source because it allows low chemical specificity, the ions produced are stable and the ionisation process is unlimited in mass (Wilm 2011). In addition, its ability to ionise samples directly from the liquid phase at atmospheric pressure makes it highly compatible with traditional chromatographic separation techniques. In ESI, a high voltage (positive in the ESI+ mode, negative in the ESI- mode) is applied to a metal capillary needle (anode/cathode for ESI+/ESI-), through which the analyte solution flows (Figure 1.30). This causes the liquid to assume a conical shape, known as the Taylor cone (Taylor 1964), and to form a mist of charged droplets

(aerosol) of the same charge as the needle. Since the droplets are charged, they repel each other and are drawn toward a counter electrode (cathode/anode for ESI+/ESI-), in the direction of the sampling orifice, losing solvent, shrinking and breaking up into smaller droplets. Continuous solvent evaporation-Coulomb fission cycles produce gas-phase ions that will travel to the vacuum chamber of the mass spectrometer (Wang and Griffiths 2008).

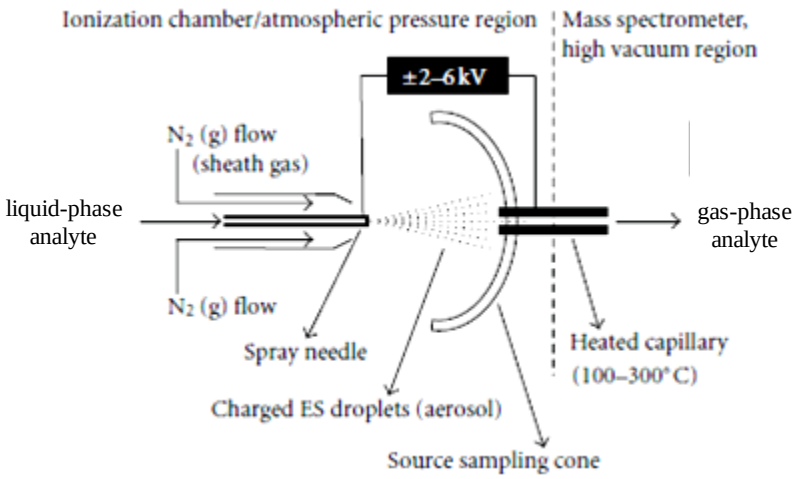


Figure 1.30 Schematic representation of an ESI ion source. (Adapted from Banerjee and Shyamalava Mazumdar 2012).

In ToF analysers (Figure 1.31), after the sample is ionised, ions are accelerated through a potential *V*, and will gain a kinetic energy, so that

$$mv^2/2 = zeV \qquad \textbf{Equation 1.14}$$

where *m* is the mass of the ion, *z* the number of charges on the ion and *e* the charge of an electron. Then:

$$m/z = 2eV/v^2 \qquad \textbf{Equation 1.15}$$

$$(m/z)^{1/2} = (2eV)^{1/2} \times t/d \qquad \textbf{Equation 1.16}$$

where *t* is the time needed for the ion to travel down a drift tube of length *d* and reach the detector. Therefore, for a given instrument operating at constant accelerating potential, *d*

and V are constants, and $t \propto (m/z)^{1/2}$, i.e., *m/z* values are determined by measuring the time that ions take to travel from the source to the detector.

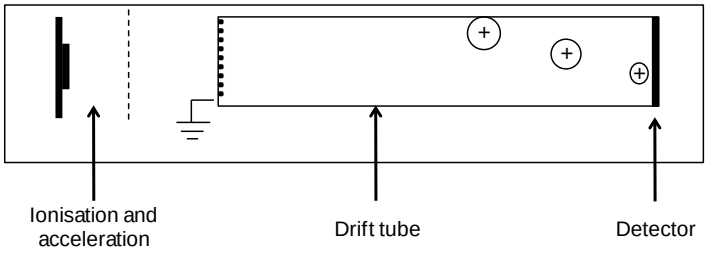


Figure 1.31 Schematic representation of a linear ToF mass spectrometer. (Adapted from Wang and Griffiths 2008).

1.3.3.2 Liquid chromatography-mass spectrometry (LC-MS)

Although samples can be directly injected into a mass spectrometer (direct infusion mass spectrometry, DIMS), MS metabolic profiling studies of biofluids commonly involve prior chromatographic separation, to reduce ion suppression resulting from many competing analytes entering the spectrometer at the same time (Want et al. 2007). Liquid chromatography (LC), which separates metabolites based on their different affinity to the liquid mobile phase and the solid stationary phase, is one of the methods most often coupled to MS. Compared to gas-chromatography (GC), it offers the advantage of allowing the analysis of metabolites which are not readily volatilised (such as nucleosides, amino acids and sugars), thus avoiding the need for complex derivatization procedures, often required in GC-MS studies. Moreover, the multiple combinations of mobile phases and separation columns make LC-MS a highly flexible method with great metabolic coverage potential (Lenz and Wilson 2007). Conventional LC-MS based on high performance liquid chromatography (HPLC) may, however, show important limitations in terms of chromatographic resolution, sensitivity and analysis times. An important upgrade to overcome these limitations has been the introduction of ultra-performance liquid chromatography – mass spectrometry (UPLC-MS) about ten years ago, which, by employing smaller diameter column packing (1.4-1.7 μm particles), smaller columns (2.1 mm i.d.), increased pressures (up to 15 kpsi) and high flow rates (up to ca. 600 $\mu L \cdot min^{-1}$), has allowed great improvements in resolution and sensitivity (Figure 1.32) (Lenz and Wilson 2007; Want et al. 2007).

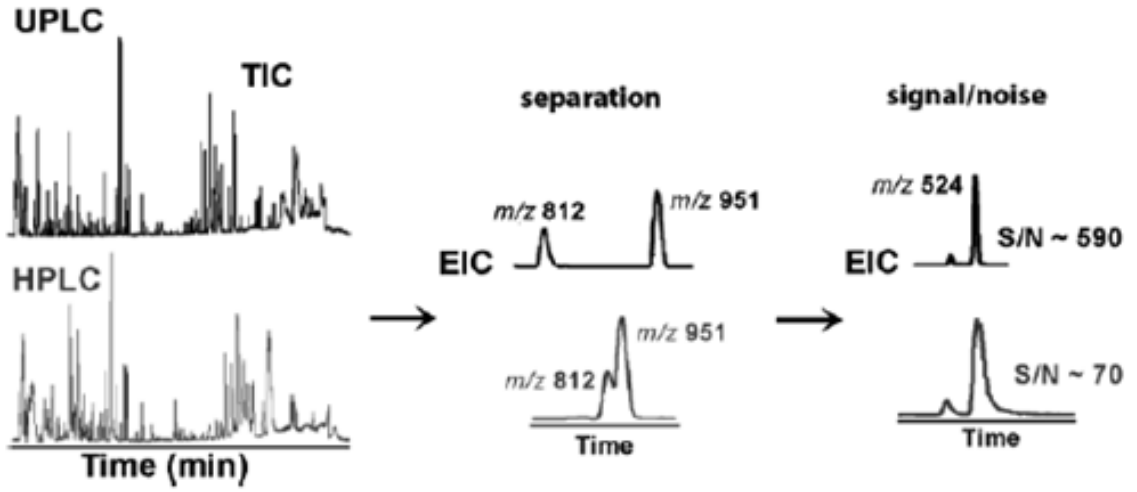


Figure 1.32 Comparison between UPLC-MS and HPLC-MS chromatograms. (Adapted from Want et al. 2007).

Depending on the biological matrix investigated and also on the main target analytes, different LC columns may be used, such as reverse phase (RP) columns for separating medium polar and nonpolar metabolites, and hydrophilic interaction chromatography (HILIC) columns for analysing highly polar metabolites which are not retained well in RP columns (Want et al. 2007). In general, UPLC-MS can provide the detection of thousands of features in a given sample, the main difficulty being the identification of those features. Indeed, extensive LC-MS databases are still lacking, hampering the fast, automated identification of metabolites, and unambiguous structural information often requires the acquisition of tandem (MS/MS) mass spectral data to obtain fragmentation patterns (Want et al. 2007).

1.3.4 Statistical tools in metabolomics

Metabolomic datasets are typically composed of several tens to hundreds of complex metabolic profiles, each comprising thousands of variables. The analysis of such a sheer data volume to extract relevant information is only possible with the help of appropriate statistical tools, such as multivariate methods. These tools are generally applied with the purpose of providing graphical overviews of data, discriminating between sample classes and highlighting the variables important for classifications. Some of these methods, namely those employed in this thesis, will be described in the following sections,

addressing also the necessary data pre-treatment steps and the validation procedures carried out to verify the robustness of multivariate models.

1.3.4.1 Data pre-treatment

Prior to applying multivariate methods, spectral datasets are typically organized into tables, where each row constitutes an *i* observation (e.g., samples spectra) and columns represent *j* variables (e.g., spectral intensities) forming an *i*×*j* matrix X. A number of pre-treatment steps may then follow: i) peak alignment, ii) normalisation and iii) scaling and/or transformation (Veselkov et al. 2011).

Peak alignment aims at minimizing shifts in the position of signals corresponding to a same metabolite, from sample to sample. In ^{1}H NMR, these shifts may originate from small differences in pH, ionic strength or temperature, as these parameters affect the effective shielding of protons, by altering the ionisation state of ionisable functional groups (e.g., carboxyl or amino), the binding to ionic metal species (e.g., citrate binding to Ca^{2+} or Mg^{2+}) or the strength of hydrogen bonds (Xiao et al. 2009). One of the approaches used to minimize NMR signal drifts is called spectral binning or bucketing and consists of dividing the spectrum into intervals of fixed/variable width (usually ranging from 0.01 to 0.05 ppm) and taking the areas of those intervals as variables. Although this allows mitigating small deviations in peak positions across samples (as well as reducing the number of variables), spectral resolution is greatly reduced, hampering the later interpretation of variations. Alternatively, peak alignment algorithms have been developed to align NMR signals without affecting spectral resolution. Examples of such algorithms include genetic algorithms (Forshed et al. 2003), recursive segment-wise peak alignment (RSPA) (Veselkov et al. 2009), correlation shifting of spectral intervals combined with fast Fourier transform (*i*coshift) (Savorani et al. 2010) or hierarchical cluster-based peak alignment (CluPA) (Vu et al. 2011). With regard to LC-MS data, shifts in retention times between samples may occur due to the slight variable interaction of the analytes with the stationary phase of the chromatographic column throughout the run. Therefore, after peak picking (the first step in the pre-treatment of LC-MS data) alignment strategies may also be employed to minimise variations in peak positions (Katajamaa et al. 2006; Smith et al. 2006).

Spectral normalization, a table row operation, is applied to compensate for differences in sample concentration or amount, allowing all samples to be directly comparable (Craig et al. 2006). There are several approaches to normalise data, the most common being normalization to total area, where each variable is divided by the total integral of the spectrum. This method assumes that the total spectral area depends only on sample dilution (concentration). However, in cases where samples contain extreme amounts of single metabolites (as is often the case in urine, due to food or xenobiotics ingestion, or in tissues containing high lipid amounts), this assumption may be wrong and cause individual peak variations to be masked. Hence, other normalization methods were developed, namely probabilistic quotient normalization (also known as median fold-change normalization) (Dieterle et al. 2006; Veselkov et al. 2011), locally weighted scatter plot smoothing (LOESS) (Veselkov et al. 2011) or group aggregating normalization (GAN) (Dong et al. 2011). The choice of the best normalization method will depend on sample type, size of dataset and on the problem to be addressed (Dieterle et al. 2006; Veselkov et al. 2011; Dong et al. 2011; Kohl et al. 2012). In this work, all data were normalized through PQN (although, in some instances, other normalization procedures were also tested). PQN is based on the calculation of the most probable dilution factor by looking at the distribution of the quotients resulting from dividing each signal intensity in a test spectrum by the corresponding intensity in a reference spectrum (usually, the median spectrum of the control group). The resulting factors are then used to normalize the data.

Another important pre-treatment step consists of scaling, a table column operation aiming at reducing the weight given by multivariate projection methods to high intensity peaks (variables) compared to smaller ones (Craig et al. 2006). This operation is performed on each variable across all samples before multivariate analysis, and commonly starts by subtracting the column mean from each value in the column (mean-centering). Afterwards, each variable is divided by a scaling factor, for instance the standard deviation of the respective column (UV – unit variance scaling) or the square root of the standard deviation of the respective column (Par – pareto scaling).

Detailed information about the pre-treatment strategies applied in this work to NMR and UPLC-MS data are present in the subchapters 2.3.5 and 2.4.3, respectively.

1.3.4.2 Multivariate analysis (MVA)

Principal Component Analysis

Principal component analysis (PCA) is a projection method used for exploratory data analysis, which does not assume any particular distribution of data, thus being classified as a non-supervised method. The principle behind PCA is to convert the original variables into a set of uncorrelated (orthogonal) new variables, called principal components (PCs, *PC1*, *PC2*, ..., *PCn*), thereby defining a lower *n*-dimensional space to represent the data. The PCs are linear combinations of the original variables (Equation 1.17, where a_{11}, a_{12}, etc. are appropriate weighing coefficients and d_1, d_2, ..., d_n the original variables), computed so that the first PC accounts for most of the variation in the dataset, the second PC accounts for the second largest variation, and so on, always obeying the constraint that all PCs are linearly orthogonal to each other.

$$PC1 = a_{11}d_1 + a_{12}d_2 + \cdots + a_{1n}d_n$$
$$PC2 = a_{21}d_1 + a_{22}d_2 + \cdots + a_{2n}d_n, etc. \qquad \textbf{Equation 1.17}$$

As a result, the original data matrix *X* can be decomposed in to a scores matrix *T* (containing information on the position of samples in the new lower-dimensional space), a loadings matrix *P* (representing the way in which the original variables contribute to the scores), and a residuals matrix *E* (part of *X* that is not explained by the model), as shown in Equation 1.18 (where *t* denotes the transpose) (Trygg et al. 2007).

$$X = TP^t + E \qquad \textbf{Equation 1.18}$$

By plotting PCA scores and loadings (Figure 1.33), a graphical overview of data is obtained, allowing groups, trends and outliers to be identified. In a scores scatter plot, each point represents an observation (sample) and the proximity between points reflects sample similarity, whereas the loadings plot shows the influence (weight) of individual *X*-variables on the observed scores distribution.

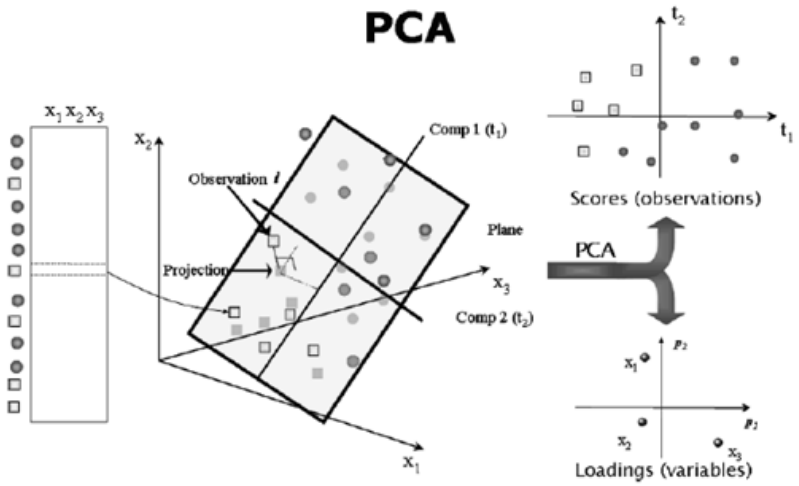


Figure 1.33 Geometrical representation of principal component analysis (PCA). (Adapted from Trygg et al. 2007).

Partial-Least Squares- and Orthogonal Projection to Latent Structures-Discriminant Analysis

Partial-least squares (PLS) is a pattern recognition method used to assess the quantitative relationship between the original data matrix *X* and a second matrix *Y* containing quantitative values (e.g., age of subjects, metabolite concentrations, etc.) (Wold et al. 2001; Trygg et al. 2007). As in PCA, the new variables defined, called latent variables (LVs), are linear combinations of the original variables; however, in PLS, LVs are iteratively obtained to model both the data matrix *X* and the correlation with *Y* (unlike PCs which only model *X*), so that the first LV is the best predictor for the information contained in the *Y* matrix. PLS models can be defined by the following equations:

$$Model\ of\ X: X = TP^t + E \qquad \textbf{Equation 1.19}$$

$$Model\ of\ Y: Y = TC^t + F \qquad \textbf{Equation 1.20}$$

where *t* denotes the transpose, *T* represents the scores matrix (common to both *X* and *Y*), *P* and *C* the loadings matrices, and *E* and *F* the residuals of *X* and *Y*, respectively (Trygg et al. 2007). The *T* scores are linear combinations of the original *X* variables and the weights *W* (quantitative relation between *X* and *Y*) and, therefore, *T* can be expressed as:

$$T = XW(P^tW)^{-1} \qquad \textbf{Equation 1.21}$$

By combining Equation 1.20 and Equation 1.21, the PLS regression can be defined by:

$$Y = XB + F \qquad \textbf{Equation 1.22}$$

where *B* contains the PLS regression coefficients defined as $B=W(P^tW)^{-1}C^t$. An important parameter for assessing variable relevance in PLS regression is the variable importance in the projection (*VIP*). *VIP* values reflect the variables' importance both with respect to *X* and *Y*, i.e., their correlation to all the responses, and can be defined as

$$VIP_{Ak} = \sqrt{\left(\sum_{a=1}^{A}\left(w_{ak}^2 \times (SSY_{a-1} - SSY_a)\right) \times \frac{K}{(SSY_0 - SSY_A)}\right)} \qquad \textbf{Equation 1.23}$$

where *SSY* is the sum of squares of the *Y* matrix for a given dimension *a*. Variables with large *VIP* (typically >1), are the most relevant for explaining *Y*. *B* coefficients and *VIP* values may be used for variable selection, as discussed ahead.

Metabolomic studies frequently aim at differentiating sample classes (e.g., control vs. disease, treated vs. non-treated, male vs. female), therefore, PLS regression is often used in combination with discriminant analysis (PLS-DA). In PLS-DA, the Y matrix contains qualitative values and all the above principles apply, the goal being to define the LVs which maximize class discrimination (Trygg et al. 2007) (Figure 1.34, left).

PLS (or PLS-DA) models are negatively affected by systematic variation in the *X* matrix that is not related to the *Y* matrix. Orthogonal projection to latent structures (OPLS) is a pre-processing method aiming at removing variation from *X* (descriptor variables) that is not correlated to *Y* (property variables) (Trygg and Wold 2002). In mathematical terms, this is equivalent to separate the systematic variation in *X* in two parts, one that is linearly related to *Y* and one that is unrelated (orthogonal) to *Y*. These are the *Y*-predictive ($T_PP_P^t$) and the *Y*-orthogonal ($T_OP_O^t$) variations; only the first is used for modelling *Y* ($T_PC_P^t$), as expressed in the equations below, where *E* and *F* are residual matrices of *X* and *Y*, respectively (Trygg et al. 2007).

$$Model\ of\ X: X = T_PP_P^t + T_OP_O^t + E \qquad \textbf{Equation 1.24}$$

$$Model\ of\ Y: Y = T_PC_P^t + F \qquad \textbf{Equation 1.25}$$

Thus, in OPLS-DA, the between-class and the within-class variations are separated, facilitating model interpretation (Figure 1.34 right). The direction separating the classes is available as a single *Y*-predictive score vector, which simplifies the interpretation of the corresponding loadings profile.

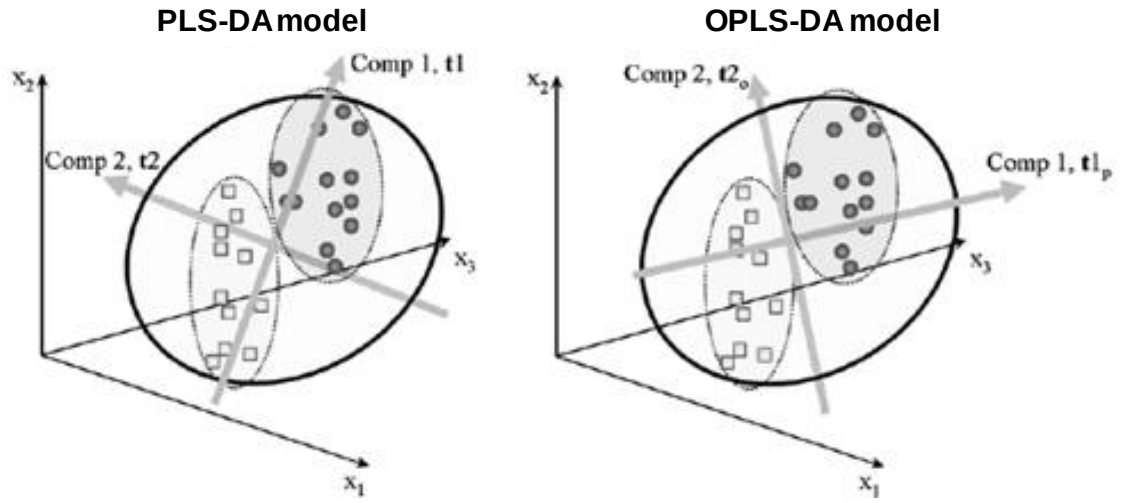


Figure 1.34 Geometrical representation of PLS-DA and OPLS-DA models. (Adapted from Trygg et al. 2007).

1.3.4.3 Variable selection

In some instances, the complex nature and large variability of metabolomic datasets may hamper the building of robust classification models. Variable selection based on multivariate methods constitutes a good approach to deal with this problem, as the removal of irrelevant, unreliable or noisy variables can reduce model complexity, improve model predictive power and facilitate the interpretation of results (Andersen and Bro 2010). Several methods of variable selection have been described in the literature, namely those based on the variable importance in the projection (VIP) values (Quintás et al. 2011; Sun et al. 2012), genetic algorithms (Cavill et al. 2009; Di Anibal et al. 2011; Lin et al. 2011), interval partial least-squares discriminant analysis (*i*-PLS) (Di Anibal et al. 2011), random forest (Lin et al. 2011; Kaur et al. 2013), among others. In this work, an in-house developed variable selection method was applied in some cases to enhance models' predictive accuracy (Diaz et al. 2013). This method is based on the intersection of three variable selection methods, which rely on the assessment of model parameters resulting

from a first PLS-DA, namely VIP values, their standard errors (VIP_{cvSE}), *b*-coefficients and their standard errors (b_{cvSE}).

1.3.4.4 Validation of classification models

The apparent separation of sample classes in a scores plot is not sufficient to attest the validity or predictive power of a classification model. In metabolomic studies, the often small number of samples compared to the high number of variables can lead to overfitting and optimistic predictive results. Hence, proper validation strategies must be applied in order to estimate how well a PLS-DA model can classify samples and predict class membership. Several approaches can be used, namely cross validation (CV) methods and permutation testing (randomisation) (Westerhuis et al. 2008). Cross validation consists of splitting the data into calibration and validation subsets. Two commonly used variations of this type of validation (used in this work) are *k*-fold (e.g., seven-fold) and Monte Carlo cross validation (MCCV). In the *k*-fold cross validation, the data are divided into *k* mutually exclusive subsets of equal size and each subset is used as training set for the remaining subsets pooled together. When *k* equals the number of samples, this method is called leave-one-out cross validation. In Monte Carlo cross validation, 40-60% of samples, randomly selected, are used as calibration set to predict the class of the remaining samples (Xu and Liang 2001). This procedure is repeated n times, typically 500-1000 times (or iterations), each time randomly changing the composition of the calibration and validation subsets. In this way, a higher number of possible combinations for dataset partitioning is performed, while for *k*-fold CV each sample is tested only once.

One of the outputs of cross validation methods is a confusion matrix expressing the difference between actual (observed) and predicted classes (Table 1.4). If the two classes compared are named ‘positive’ and ‘negative’ (e.g., for designating disease and control classes), the confusion matrix will show the number of ‘positive’ and ‘negative’ samples correctly classified (true positives and true negatives, respectively), as well as the number of misclassified samples (false positives and false negatives), thus allowing the determination of model sensitivity, specificity and classification rate (Fawcett 2006).

Table 1.4 Confusion matrix obtained by CV of two classes.

	Observed	
Predicted	**Positive**	**Negative**
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

Sensitivity (also called true positive rate) expresses the proportion of true positives correctly identified as such (Equation 1.26), while specificity (or true negative rate) expresses the proportion of negatives correctly classified (Equation 1.27). The classification rate expresses the overall accuracy, i.e. the proportion of samples correctly classified (Equation 1.28).

$$sensitivity\ (TPR) = \frac{TP}{TP + FN} \qquad \textbf{Equation 1.26}$$

$$specificity\ (1 - FPR) = \frac{TN}{FP + TN} \qquad \textbf{Equation 1.27}$$

$$classification\ rate\ (CR) = \frac{TP + TN}{TP + FN + FP + TN} \qquad \textbf{Equation 1.28}$$

The sensitivity and 1-specificity results can then be plotted as coordinates in a receiver operating characteristic (ROC) space plot (Figure 1.35). A perfect classification model would have both sensitivity and specificity equal to 1, while one with no predictive ability would have sensitivity equal to 1-specificity (Fawcett 2006).

Another useful parameter to evaluate the robustness of a classification model is the Q^2 value, which expresses the model's predictive power, and is given by

$$Q^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2} \qquad \textbf{Equation 1.29}$$

where $\hat{y}_i$ refers to the predictive value of class membership for sample i and $\overline{y}$ refers to the mean value of y for all samples (Westerhuis et al. 2008). The closer Q^2 is to 1, the higher is

the model's predictive power. An optimal Q^2 of 1 is difficult to obtain as it will depend not only on inter-class variability, but also on intra-class variability, making it difficult to define a general Q^2 value that corresponds to a good discrimination. One way to assess how good Q^2 values are consists of comparing their distributions in original and permuted models (Westerhuis et al. 2008).

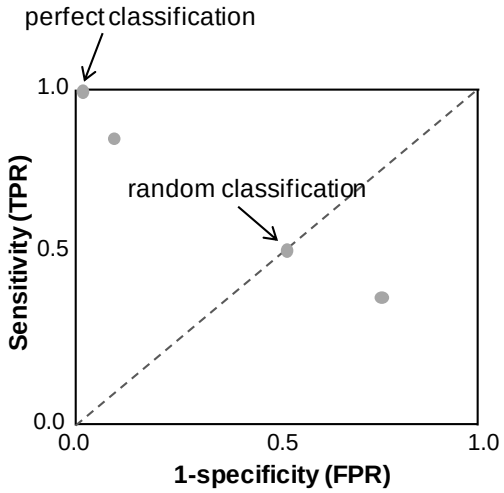


Figure 1.35 ROC space plot illustrating a perfect classification (TPR=1 and FPR=0) and a random classification (TPR=FPR). Models falling above the dashed line classify better than those falling below.

Permutation testing consists of randomly changing class membership (*Y* matrix), leaving the *X* matrix intact, and repeating the cross validation procedures mentioned above. The model with permuted classes is expected to clearly show lower classification accuracy and predictive power (compared to the model with true classes assigned), as the randomly formed groups are not expected to show consistent differences between them (Westerhuis et al. 2008).

Ultimate validation of a classification model can be performed by predicting class membership of independent samples from a plausible related population. This type of validation is called external validation (Bleeker et al. 2003; Steyerberg et al. 2003). Although generally less optimistic (because classification models tend to perform better on data used for model building), external validation resembles a more realistic scenario in which a classification model is used for diagnostics or screening. Therefore, it should be performed whenever the experimental setup (e.g., the number of samples available) allows it.

1.3.4.5 Univariate statistics

In addition to multivariate analysis, univariate calculations are often included in metabolomic studies to assess the relevance of a given variable (metabolite). Unlike MVA, classical univariate analysis is applied to one variable at the time, and includes, e.g., significance tests (*t*-test, Wilcoxon test, Mann-Whitney test) and analysis of variance (ANOVA) (Sokal and Rohlf 2012). Although the need of applying univariate analysis after performing MVA is subject of discussion, a recent paper on this matter concluded that both methodologies give complementary information, as long as they are interpreted within the corresponding statistical framework (uni or multivariate) (Saccenti et al. 2013). In the context of this thesis, the univariate Wilcoxon signed rank test (paired samples) and the Wilcoxon rank sum test (unpaired samples) were performed to statistically evaluate the difference in specific metabolite levels between two groups.

While *p*-values report the statistical significance of variations, they do not provide any information about the magnitude of the effect of interest or its precision, besides being influenced by sample size (Nakagawa and Cuthill 2007). Therefore, to address these issues, another parameter has been suggested to be included in biological studies: the effect size (*d*). In addition to providing information about the direction and strength of the relationship between variables, effect size estimates are not sensitive to sample size, providing a common metric to compare results across studies (Berben et al. 2012). This parameter and respective confidence interval for 95% confidence may be calculated according to Equation 1.30, Equation 1.31 and Equation 1.32, where $\bar{x}_1$ and $\bar{x}_2$ are the averages, s_1 and s_2 are the standard deviations, and n_1 and n_2 are the number of samples of groups 1 and 2, respectively.

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s} \qquad \textbf{Equation 1.30}$$

$$s^2 = \frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2}{n_1 + n_2 - 2} \qquad \textbf{Equation 1.31}$$

$$d \pm 1.96 \sqrt{\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}} \qquad \textbf{Equation 1.32}$$

1.3.4.6 Correlation analysis

Assessing the statistical correlation between two variables (metabolites) can be useful to assist the biochemical interpretation of their variations. For normally distributed variables (*x* and *y*, forming *n* pairs) the Pearson correlation coefficient, *r*, estimates the degree of linear association between variables and is given by:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \qquad \textbf{Equation 1.33}$$

When the variables are not normally distributed or the relationship between them is not linear, the measurement of Spearman rank correlation coefficient, r_S, may be more appropriate (Zou et al. 2003). Correlation coefficients vary between -1 (perfectly negative association between variables) and 1 (perfectly positive association between variables), r=0 meaning no association at all.

Correlation analysis is also at the basis of a method developed primarily to help the identification of metabolites in NMR metabolomic studies, the so-called Statistical Total Correlation Spectroscopy (STOCSY) (Cloarec et al. 2005). STOCSY is based on the properties of the correlation matrix *C*, computed from a set of sample spectra according to Equation 1.34,

$$C = \frac{1}{n-1} X_1^t X_2 \qquad \textbf{Equation 1.34}$$

where X_1 and X_2 denote the experimental matrices and *C* is the correlation matrix where each value represents the correlation coefficient between two variables of the matrices X_1 and X_2 (Cloarec et al. 2005). In general, STOCSY takes advantage of the multicollinearity of the intensity variables in a set of spectra to generate a plot where highly correlated signals are highlighted, thus providing information on molecular or biochemical relationships. This can be applied to data obtained from a single NMR experiment type (^{1}H NMR, Figure 1.36a), as well as to data generated by the NMR observation of different nuclei (e.g., ^{1}H and ^{31}P NMR, designated heteronuclear STOCSY, HET-STOCSY) (Wang et al. 2008). A further development of this method consists of applying it to data obtained

by different analytical techniques, for instance ^{1}H NMR and MS (Figure 1.36b) (Crockford et al. 2006), or ^{1}H NMR and mid-infrared (MIR) (Graça et al. 2013). This multispectroscopic approach is designated statistical heterospectroscopy (SHY) (Crockford et al. 2006).

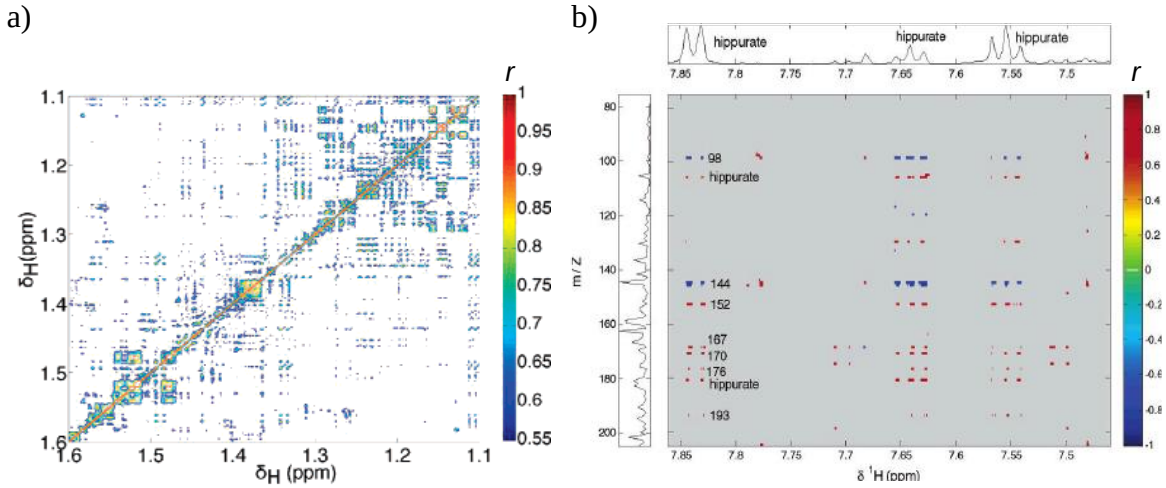


Figure 1.36 a) ^{1}H-^{1}H NMR STOCSY map (adapted from Keun et al. 2008) and b) UPLC-MS – ^{1}H NMR SHY map showing positive correlations of hippurate in both spectroscopic domains. (Adapted from Crockford et al. 2008).

1.3.5 Metabolomics of lung cancer: state of the art

The application of metabolomics to the study of cancer has intensified over the years, as seen by the exponential increase in the number of papers published in the field (Figure 1.37a). Regarding the biological matrix used, the first studies were based on the analysis of tissue samples, while biofluid analysis has been increasingly performed over the last ten years (Figure 1.37a), blood serum/plasma and urine being the most studied biofluids. In terms of the analytical techniques employed, NMR has been preferred for tissues, likely in relation with the possibility of direct analysis by HRMAS, whereas biofluids have been mostly studied by MS methods, taking advantage of their generally higher sensitivity (Figure 1.37b).

Concerning the types of cancer studied, breast, colorectal, lung, brain and prostate cancers have been the main focus in metabolomics, constituting over 60% of the published papers in this area (Figure 1.37c). These numbers likely reflect cancer incidence, as breast, colorectal and lung cancers are those with highest prevalence worldwide (Figure 1.1). Still,

other less frequently cancer types have also been investigated namely hepatocellular and renal cell carcinomas, ovarian and cervical, gastric and oesophageal cancers, among others.

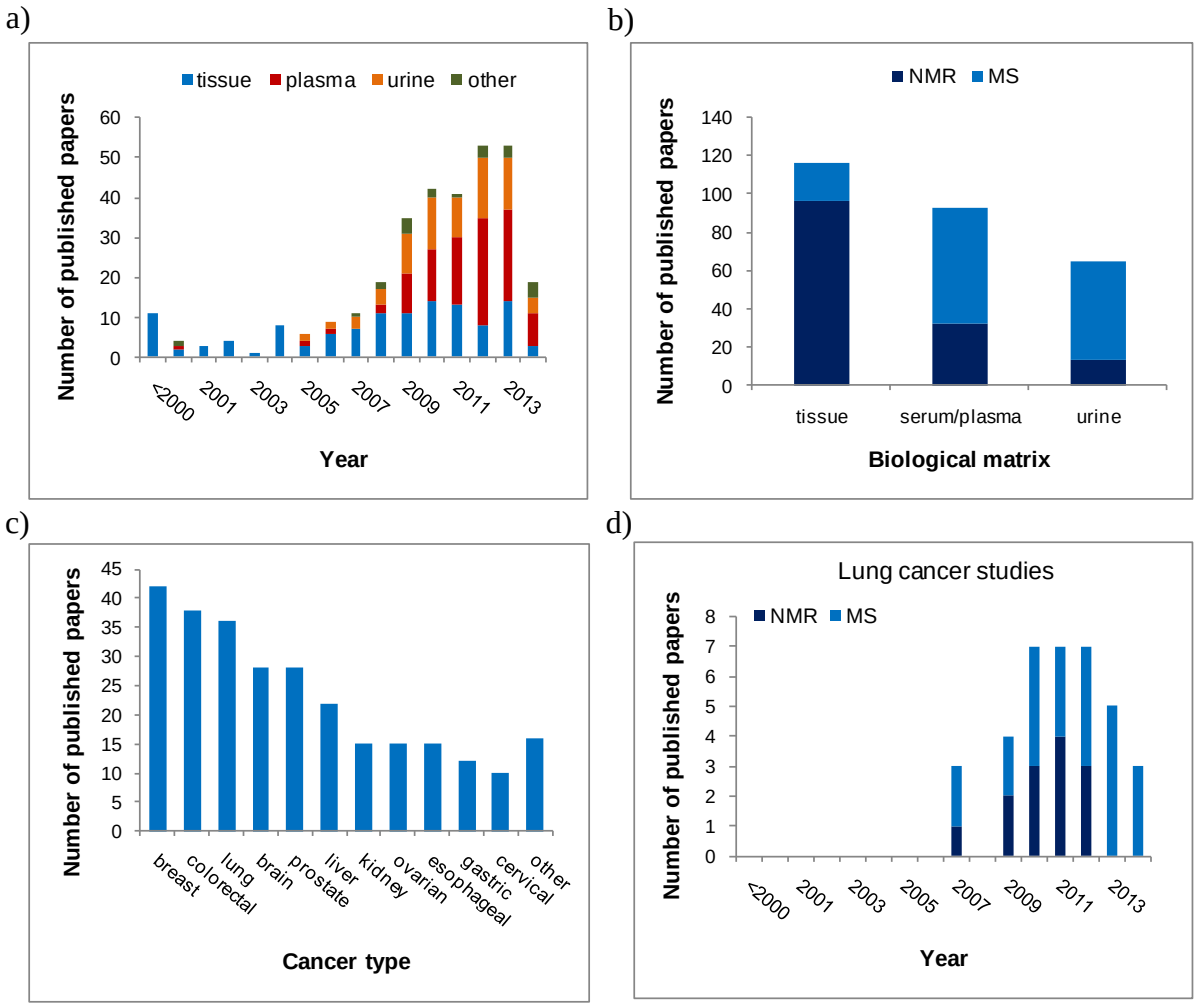


Figure 1.37 Estimate number of published papers on metabolomics applied to human cancer studies distributed by a) year and biological matrix (other matrices include bile, saliva, exhaled breath condensate, cerebrospinal fluid, pleural effusion, seminal fluid and prostatic fluid), b) biological matrix and analytical platform used, c) type of cancer studied (other types of cancer include thyroid, myeloma, liposarcoma and osteosarcoma). d) Number of papers published on clinical metabolomic studies of human lung cancer. Source: Web of Science database (all years, up to June 2014).

The metabolomic study of lung cancer has grown considerably over the last five years (Figure 1.37d). Indeed, when the research presented in this thesis was initiated, the information on the metabolic behaviour of lung tumours and on their signatures in biofluids was very limited, thus justifying the relevance and novelty of the work conducted. In this section, the studies using metabolomics to characterize human lung cancer will be reviewed, starting by the analysis of tissues (or their extracts) and presenting

thereafter the most important findings emerging from the analysis of biofluids (blood plasma or serum, urine and others), using either targeted profiling of specific compound families (earlier studies) or global untargeted metabolomics (studies performed over the last five years). A table listing all the studies mentioned in this section, with details on the sample numbers, the analytical techniques and the multivariate analysis methods used, is presented in Annex III.

Earlier metabolic studies of lung tumours were based on the NMR analysis of tissue extracts to measure differences in the levels of specific metabolites (e.g., lactate, creatine, choline) and assess their diagnostic and prognostic value (Hanaoka et al. 1993; Guo et al. 2004; Yokota et al. 2007). In another study using NMR and GC-MS stable isotope resolved metabolomics (SIRM), ^{13}C enrichment of particular metabolites (lactate, succinate, citrate, alanine, glutamate, aspartate), after infusion of uniformly labelled ^{13}C-glucose into patients, suggested more active glycolysis and TCA cycle in tumour tissues (Fan et al. 2009). More recently, the composition of lung tissue extracts has been more comprehensively investigated by capillary electrophoresis (CE)-ToF MS and the levels of several amino acids, adenylates and guanylates, lactate and other organic acids were found to be increased in tumours compared to surrounding normal tissue (Kami et al. 2013). However, the low number of patients considered (n 9) precludes solid conclusions from this study. Moreover, metabolic profiling of tissue extracts has some important limitations, such as the potential loss or modification of cellular components, and the selectivity of the extraction procedures towards more hydrophilic or lipophilic compounds, which may limit the information about the metabolic events taking place *in vivo*.

High-resolution angle magic spinning (HRMAS) NMR spectroscopy, on the other hand, enables the direct analysis of intact tissue biopsies, having been increasingly applied to characterize the metabolic profile of tumour tissues (Sitter et al. 2009; Moestue et al. 2011). In regard to lung cancer, the first HRMAS NMR studies of human lung tumours thoroughly describing the tissue metabolome and the differences between tumour and non-involved adjacent (control) tissues have been published by our group (Rocha et al. 2010; Duarte et al. 2010). Although involving relatively low number of patients (twelve in the first study, twenty four in the second), these papers provided consistent evidence of distinct metabolic profiles for tumour and control tissues, mainly based on the levels of glucose, lactate, choline-containing compounds, acetate and inositols. A subsequent study by Chen

et al. corroborated several of these findings and showed that some metabolite levels (lipids, lactate, glucose, aspartate and choline metabolites) depended on the site of tissue collection (centre or periphery of the tumour) (Chen et al. 2011). Also, the above mentioned studies suggested a dependency of the tissue metabolic profiles relatively to the tumour histological type. This dependency was also explored by Jordan et al., who attempted the distinction between adenocarcinomas (AdC) and squamous cell carcinomas (SqCC), based on the correlation between tissue and serum profiles (from fourteen patients). However, no detailed information on possible metabolic markers was advanced in that study (Jordan et al. 2010). More recently, resulting from the work presented in Chapter 3 of this thesis, our group has reported the analysis of a considerably large sample set (over one hundred tissue samples from fifty six patients) (Rocha et al. 2014), further demonstrating the potential of NMR metabolomics for the differential classification of AdC and SqCC.

Up to 2010, the studies addressing the metabolic composition of biofluids in the context of lung cancer were focused on the targeted analysis of specific compounds. In the case of blood, circulating plasma free amino acids (PFAA) have been the main targeted family investigated, based on the early studies (reviewed in Lai et al. 2005) suggesting a link between abnormal PFAA composition and cancer-induced alteration in protein metabolism. In particular, a Japanese research group has reported consistenly altered PFAA profiles in lung cancer patients, using HPLC-MS analysis, and proposed a diagnostic index based on PFAA concentrations (known as the 'AminoIndex' technology) (Maeda et al. 2010; Miyagi et al. 2011; Shingyoji et al. 2013). Blood phospholipids were another class of compounds targeted for the discrimination between lung cancer patients and healthy controls. Particularly, lysophosphatidylcholines (sn-1 lysoPC 16:0, sn-2 lysoPC 16:0, sn-1 lysoPC 18:0, sn-1 lysoPC 18:1 and sn-1 lysoPC 18:2) identified by UPLC-ToFMS were found to be decreased in cancer patients (Dong et al. 2010), whereas in a different study using direct infusion (DI)-Fourier transform ion cyclotron resonance (FTICR)-MS, oleamides, long chain acylcarnitines, LPC (18:1), LPC (20:4), LPC (20:3), LPC (22:6) and SM (16:0/1) were responsible for the discrimination between the two groups (Guo et al. 2012).

Regarding the untargeted analysis of blood serum/plasma, a study published in 2010 has reported the potential of NMR of serum to differentiate between two lung cancer histological types and also from controls (Jordan et al. 2010). However, sample numbers

were very limited (fourteen patients and seven controls) and no information was provided on putative metabolic differences between groups. Subsequently, our group has reported a more detailed NMR study of blood plasma from a significantly larger dataset (eighty five patients and seventy eight controls) (Rocha et al. 2011). As thoroughly described in Chapter 4 of this thesis (for an even larger number of subjects), this analysis has enable patients to be discriminated from healthy controls with sensitivity and specificity levels around 90%, the main metabolic differences relating to lipoproteins, organic acids, glucose and several amino acids.

Since then, a number of metabolomic studies addressing plasma/serum composition have been published, all based on MS methods. In the work by Hori and co-workers, based on GC-MS, metabolites such as TCA cycle intermediates, amino acids and their derivatives and fatty acids or related compounds, were found to significantly vary between patients and controls (Hori et al. 2011). In a different work, based on UPLC-MS, nineteen ions were suggested as potential biomarkers for lung cancer, but only two of them were identified: *p*-hydroxyphenyllactic acid and proline betaine (Cai et al. 2011). Another important outcome of this work was the difference encountered between patients at different stages of radiotherapy treatment, suggesting that the proposed methodology could be useful in the evaluation of treatment effects. In addition, Lokhov et al. published two reports focused on the analysis of deproteinized blood plasma of lung cancer patients and age-matched controls by DI-ESIMS. In the first work, a model with a classification accuracy of 92-94% was obtained, regardless of disease stage (Lokhov et al. 2012), and in the second work, addressing the same dataset, seventy metabolite ions were found to be strongly associated with lung cancer (Lokhov et al. 2013). In an interesting study, using different UPLC-MS platforms, urea, bilirubin and 3-hydroxybutanoic acid were reported to be negatively correlated with survival time of SCLC patients, suggesting elevated levels of these metabolites to be present in the blood of subjects with a poorer prognosis (Vaughan et al. 2012). Finally, in a recent study using both GC- and LC-MS, plasma from stage I lung adenocarcinoma patients and healthy controls was compared (Wen et al. 2013). Results from OPLS-DA modelling afforded thirty seven metabolites (related to amino acids, lipids and fatty acids metabolism) to contribute for class differentiation.

In the case of urine, a first study reported the targeted analysis of modified nucleosides and ribosylated metabolites from patients with different types of cancer,

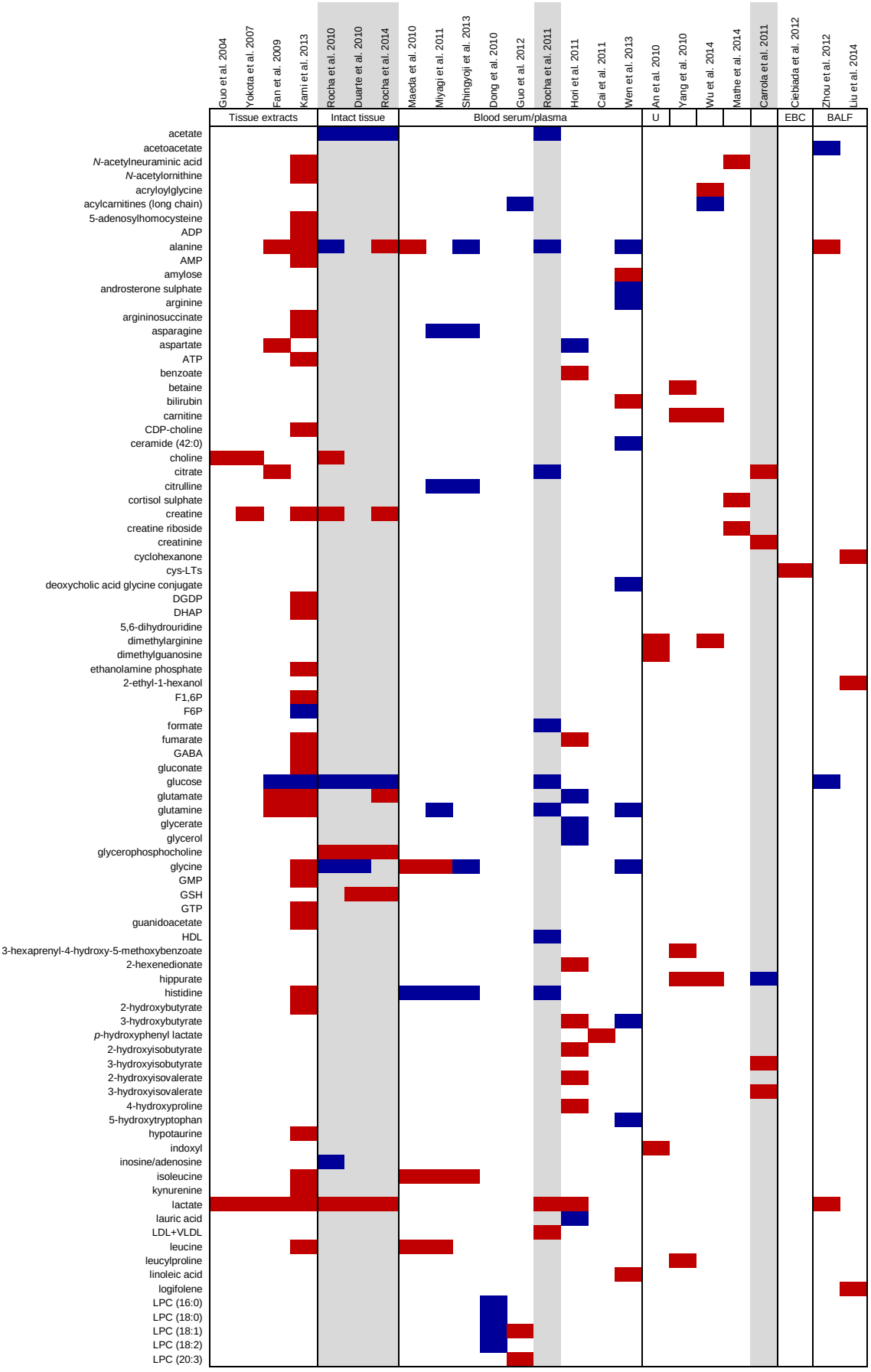
including lung cancer (Bullinger et al. 2008). Two subsequent studies, published in 2010, focused on the development of improved LC-MS methods for the untargeted profiling of urine, either by using an integrated ionization approach (An et al. 2010) or a home-devised system based on HILIC/RPLC-MS (Yang et al. 2010). Each study reported the identification of eleven candidate biomarkers, including aromatic amino acids and modified nucleosides in one case (An et al. 2010) and mainly amino acids in the other case (Yang et al. 2010). Very recently, UPLC-MS has been applied to the urinary profiling of a relatively small set of subjects (twenty controls and twenty patients), whereby twenty discriminant metabolites (including carnitine and acylcarnitines) were identified (Wu et al. 2014). Moreover, another untargeted LC-MS study comprising a much larger population (over four hundred patients and five hundred controls) highlighted creatine riboside and *N*-acetylneuraminic acid as being associated with early lung cancer diagnosis and worse prognosis (Mathe et al. 2014).

In 2011, our group has published the first untargeted NMR study of urine in the context of lung cancer (Carrola et al. 2011). At the time, the results referred to about one hundred and twenty subjects (a number that has almost doubled up to the completion of this thesis) and showed the potential of NMR metabolomics of urine to discriminate between early stage cancer patients and healthy controls. The possible confounding influence of factors like gender and age have also been modelled and found to have much lower predictive power than the presence of the disease. In other subsequent works, NMR urinary profiles have been used to attempt the prediction of cancer-associated skeletal muscle wasting (Eisner et al. 2011), and of variations in lean muscle and fat mass, in patients with advanced cancer (Stretch et al. 2012).

Other biofluids have also been investigated in the context of lung cancer, namely bronchoalveolar lavage fluid (BALF), exhaled breath condensate (EBC) and pleural effusions. A report addressing the composition of BALF and EBC samples from patients with primary lung cancer focused on a specific group of compounds, eicosanoids, assessed by enzymatic assays (Ciebiada et al. 2012). Patients presented higher levels of cysteinyl leukotrienes and leukotriene B4 in both biofluids leading the authors to propose the analysis of these compounds, known to be related to lung carcinogenesis, as a complementary detection and monitoring method. In another work, malignant and benign pleural effusions from patients with lung cancer and metastasis to the pleura were analysed

by ^{1}H NMR (Zhou et al. 2012). By applying multivariate analysis, malignant effusions could be differentiated from benign pleural effusions, based on metabolites like lactate, acetoacetate, trimethylamine-*N*-oxide and glucose. The low molecular weight volatile organic fraction of pleural effusions has also been assessed by headspace-solid phase microextraction (HS-SPME) GC-MS (Liu et al. 2014). The authors highlighted the levels of cyclohexanone, 2-ethyl-1-hexanol, 2-phenyl-2-propanol, 1,2,4,5-tetramethylbenzene and longifolene to be significantly different between benign and malignant pleural effusions.

Overall, by compiling the biochemical information provided in the studies mentioned above, it is found that over 150 metabolites, detected in several biological matrices by different metabolomic analytical platforms, have been related to lung cancer altered metabolism. Figure 1.38 summarises this information in the form of a heatmap where blue and red colours represent, respectively, decreases and increases in either lung tumour tissues or lung cancer patients in relation to control pulmonary tissues or healthy subjects. Several interesting observations emerge from this Figure. Firstly, it is seen that some changes are consistent across studies (e.g., increased lactate and decreased glucose), while others show large variability (e.g. some amino acids like alanine, glycine or tyrosine). This variability may be a reflection of several factors, including differences in the population groups studied (e.g. genetic background, lifestyle/dietary habits), but also differences in sample handling and analysis procedures. Secondly, the large majority of variations was detected in a single biological matrix, thus showing the importance of analysing different sample types in order to achieve wider metabolome coverage and get a more complete picture of metabolic events. Finally, it is important to recall that many of these studies were performed on low sample numbers (as it is clear from Table A4 in Annex III) and that many of the variations listed were not quantitatively measured or statistically validated, thus justifying the need for in-depth, rigorous studies of lung cancer's metabolic signature, as it has been attempted in this thesis.


Guo et al. 2004
Yokota et al. 2007
Fan et al. 2009
Kami et al. 2013
Rocha et al. 2010
Duarte et al. 2010
Rocha et al. 2014
Maeda et al. 2010
Miyagi et al. 2011
Shingyoji et al. 2013
Dong et al. 2010
Guo et al. 2012
Rocha et al. 2011
Hori et al. 2011
Cai et al. 2011
Wen et al. 2013
An et al. 2010
Yang et al. 2010
Wu et al. 2014
Mathe et al. 2014
Carrola et al. 2011
Ciebiada et al. 2012
Zhou et al. 2012
Liu et al. 2014
Tissue extracts
Intact tissue
Blood serum/plasma
U
EBC
BALF
acetate
acetoacetate
N-acetylneuraminic acid
N-acetylornithine
acryloylglycine
acylcarnitines (long chain)
5-adenosylhomocysteine
ADP
alanine
AMP
amylose
androsterone sulphate
arginine
argininosuccinate
asparagine
aspartate
ATP
benzoate
betaine
bilirubin
carnitine
CDP-choline
ceramide (42:0)
choline
citrate
citrulline
cortisol sulphate
creatine
creatine riboside
creatinine
cyclohexanone
cys-LTs
deoxycholic acid glycine conjugate
DGDP
DHAP
5,6-dihydrouridine
dimethylarginine
dimethylguanosine
ethanolamine phosphate
2-ethyl-1-hexanol
F1,6P
F6P
formate
fumarate
GABA
gluconate
glucose
glutamate
glutamine
glycerate
glycerol
glycerophosphocholine
glycine
GMP
GSH
GTP
guanidoacetate
HDL
3-hexaprenyl-4-hydroxy-5-methoxybenzoate
2-hexenedionate
hippurate
histidine
2-hydroxybutyrate
3-hydroxybutyrate
p-hydroxyphenyl lactate
2-hydroxyisobutyrate
3-hydroxyisobutyrate
2-hydroxyisovalerate
3-hydroxyisovalerate
4-hydroxyproline
5-hydroxytryptophan
hypotaurine
indoxyl
inosine/adenosine
isoleucine
kynurenine
lactate
lauric acid
LDL+VLDL
leucine
leucylproline
linoleic acid
logifolene
LPC (16:0)
LPC (18:0)
LPC (18:1)
LPC (18:2)
LPC (20:3)

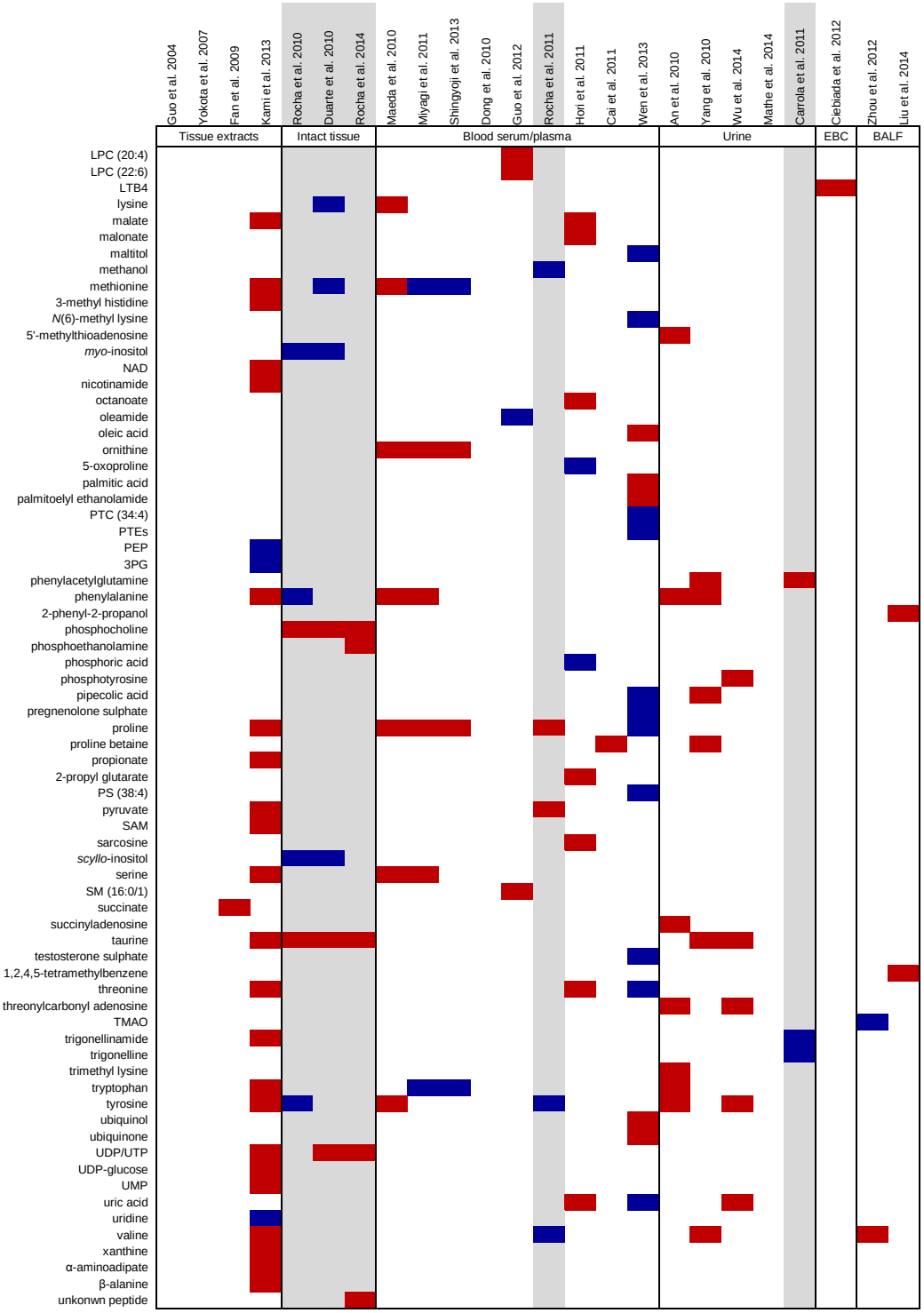


Figure 1.38 Overview of metabolites found to be decreased (in blue) or increased (in red) in either lung tumour tissues or lung cancer patients in relation to control pulmonary tissues or healthy subjects, based on a literature survey of metabolomic studies of lung cancer, including the reports derived from this thesis (highlighted in grey). Metabolites are listed in alphabetic order and literature references are given in columns, organized by biological matrices - tissue extract, intact tissue, blood serum/plasma, urine, exhaled breath condensate (EBC) and bronchoalveolar lavage fluid (BALF). (ADP: adenosine diphosphate; AMP: adenosine monophosphate; ATP: adenosine triphosphate; CDP: cytidine diphosphate; cys-LTs: cysteinyl leukotrienes; DGDP: deoxyguanosine diphosphate; DHAP: dehydroxyacetone phosphate; F1,6P: fructose 1,6-biphosphate; F6P: fructose

6-phosphate; GABA: gamma-aminobutyric acid; GMP: guanosine monophosphate; GSH: reduced glutathione; GTP: guanosine triphosphate; HDL: high-density lipoprotein; LDL+VLDL: low/very low-density lipoprotein; LPC: lysophosphatidylcholine; LTB4: leukotriene B4; NAD: nicotinamide adenine dinucleotide; PEP: phosphoenolpyruvate; 3PG: 3-phosphoglycerate; PS: phosphatidylserine; PTC: phosphatidylcholine; PTEs: phosphatidylethanolamines; SAM: *S*-adenosyl methionine; SM: sphingomyelin; TMAO: trimethylamine-*N*-oxide; UDP/UTP: uridine di/triphosphate; UMP: uridine monophosphate).

1.4 Scope and aims of this thesis

This work aimed at characterizing the metabolic alterations related to lung cancer, with a view to gain new insights into tumour metabolism and to identify metabolic markers (or profiles) with potential value in disease screening, diagnosis or treatment. Specific objectives were:

i. To characterize the metabolic differences between lung tumour tissues and adjacent non-involved pulmonary parenchyma, and assess their potential for discriminating between the two tissue types. The biochemical information derived from this study, not available from conventional histopathological analysis, is expected to provide important leads for future research on novel therapeutic targets or imaging tracers, and will eventually afford additional criteria for improving clinical decisions;

ii. To correlate tissue metabolic profiles with histological parameters, assessing, in particular, the metabolic behaviour of different tumour subtypes, with emphasis on the most frequent malignant epithelial tumours: adenocarcinoma and squamous cell carcinoma. Establishment of such correlations could be relevant in the context of lung tumours' differential diagnosis;

iii. To identify putative cancer-related metabolic alterations in easily accessible biofluids (blood plasma and urine) and explore their potential for discriminating lung cancer patients from healthy controls;

iv. To assess the possible confounding influence of the inter-individual variability arising from demographic or lifestyle factors (e.g., gender, age, smoking habits) on the predictive ability of biofluid-based classification models;

v. To identify a systemic metabolic signature for lung cancer, detectable in biofluids, and characterize its dependence on tumour histological type and stage, the ultimate goal being the development of adjunct, minimally invasive screening and/or diagnostic methods that would improve lung cancer management.

2 MATERIALS AND METHODS

This work has been carried out in the framework of a collaboration with the Faculty of Medicine (FMUC) and the University Hospitals of Coimbra (HUC, EPE), where patients were recruited. Collaboration with a local company (INDASA S.A.) was also established to recruit healthy volunteers, who integrated the control group. Support from FCT was granted through the research project FCT/PTDC/QUI/68017/2006 (January 2008 – June 2011) and the PhD grant SFRH/BD/63430/2009. Moreover, this work has been supported by CIMAGO (FMUC, University of Coimbra) and by 'Liga Portuguesa Contra o Cancro'.

2.1 Subjects

In total, 120 lung cancer patients (39 female, 81 male, average age 62 and median age 63, age range 30-81) were included in this study, following protocol approval by the Ethics Committee of the University Hospitals of Coimbra (HUC, EPE) and informed consent. All patients were diagnosed with primary malignant epithelial tumours of the lung and were hospitalized for tumour surgical removal. Final diagnosis and staging were established by histopathological evaluation, according to the 2004 World Health Organisation (WHO) classification of lung tumours (Travis et al. 2004). The histological types included were adenocarcinoma (n 50), squamous cell carcinoma (n 29), carcinoid tumour (n 15), sarcomatoid (n 9), large cell carcinoma (n 7), adenosquamous carcinoma (n 5) and small cell carcinoma (n 5). Based on the TNM staging system, tumours were classified as stage I (n 69), stage II (n 26) and stage III (n 13). For the remaining 12 cases, staging information was inconclusive or not available. In all cases, the surrounding parenchyma (non-involved tissue) presented histological characteristics of 'smoking lung' (emphysema, inflammation, desquamation), indicating exposure to cigarette smoking (although information on actual smoking habits was not available for the large majority of patients). None of the patients had received radiation or chemotherapy treatment or shared any common medication regimen previously to sample collection, although casual intake of other medicines (e.g., analgesics, antihypertensive drugs) could not be ruled out. The

complete information on the demographic and histopathological data of lung cancer patients can be found in Table A5 of Annex IV.

Regarding the control group, 95 healthy volunteers (46 female, 49 male, average age 42, median age 45, age range 22-60) were recruited within occupational medicine routine exams (at the company INDASA S.A.), after informed and signed consent. Subjects were included in the control group on the basis of a physician's assessment of their general health status (normal values in blood and urine standard clinical tests, as well as absence of major illness and chronic medication). Subjects with chronic diseases, namely diabetes, dyslipidemia and other lung diseases (e.g., COPD) were not included. The complete demographic information of control subjects can be consulted in Table A7 of Annex V.

2.2 Sample collection

The time span of sample collection and analysis was four years, from 2008 to 2012. All patients provided blood and urine samples prior to surgery, while pulmonary tissues were only collected for a subset of patients for which redundant tissue (not needed for diagnosis) was available. Sample numbers collected and analysed for each group and sample type, together with general demographic and histopathological information, are summarised in Table 2.1.

2.2.1 Lung tissues

Tumour and non-involved adjacent control tissue samples were retrieved from pulmonary surgical specimens by an experienced pathologist, within a maximum of 30 min after tumour resection to minimize ischemia-related changes. Collected tissues were then immediately snap-frozen in liquid nitrogen to stop enzymatic/chemical reactions, and stored at -80 ºC at HUC, EPE (Coimbra), before being transferred in dry ice to CICECO, Department of Chemistry at the University of Aveiro, where they were stored at -80 ºC until NMR analysis. The total storage period did not exceed six months, varying between one to three months for the majority of samples. Sampling for diagnosis was routinely performed and mirror sections of the samples used for NMR analysis were microscopically

Table 2.1 Summary of sample types and numbers analysed by the different analytical techniques used in this work.

Analytical technique	**^{1}H HRMAS NMR**	**^{1}H NMR**				**UPLC-MS**	
Sample type and group	**Tissue**	**Blood plasma**		**Urine**		**Urine**	
	Patients[a]	**Controls**	**Patients**	**Controls**	**Patients**	**Controls**	**Patients**
No. of subjects	56	94	106	91	109	50	48
Gender[b]							
Male	40 (71%)	49 (52%)	76 (72 %)	48 (53%)	77 (71%)	25 (50%)	37 (77%)
Female	16 (29%)	45 (48%)	30 (28%)	43 (47%)	32 (29%)	25 (50%)	11 (23%)
Average; median age (range)	62; 64 (36-81)	43; 44 (22-60)	62; 63 (30-81)	42; 45 (22-59)	63; 63 (30-81)	49; 49 (38-59)	61; 61 (30-84)
Smoking habits[b]							
Smokers		29 (31%)		28 (30%)		14 (28%)	
Non-smokers	'smoking lung'[c]	44 (47%)	'smoking lung' [c]	43 (47%)	'smoking lung' [c]	22 (44%)	'smoking lung' [c]
Ex-smokers		21 (22%)		20 (23%)		14 (28%)	
Histological type[b]							
Adenocarcinoma	19 (34%)		40 (38%)		42 (38%)		22 (46%)
Squamous cell carcinoma	19 (34%)		27 (25%)		27 (24%)		11 (23%)
Carcinoid tumour	4 (7%)		15 (14%)		15 (14%)		7 (15%)
Sarcomatoid carcinoma	6 (11%)	-	8 (8%)	-	8 (7%)	-	3 (6%)
Large cell carcinoma	4 (7%)		7 (7%)		7 (6%)		4 (8%)
Small cell carcinoma	2 (4%)		5 (5%)		5 (5%)		1 (2%)
Adenosquamous carcinoma	2 (4%)		4 (4%)		7 (6%)		0 (0%)
TNM stage[b]							
I	29 (54%)		61 (64%)		63 (65%)		31 (72%)
II	19 (35%)	-	24 (25%)	-	23 (24%)	-	5 (12%)
III	6 (11%)		10 (11%)		11 (11%)		7 (16%)

[a] Paired tumour and non-involved (control) tissues were collected for each patient. For one patient, the control sample was not available, while for other two patients there were two tumour tissue fragments, giving a total of 115 tissue samples analysed. [b] Percentages were calculated within each group. [c] Pulmonary parenchyma with histological characteristics indicating exposure to cigarette smoke; smoking history not available.

observed so that the percentages of tumour cells, necrosis and inflammatory tissue could be determined.

2.2.2 Biofluids: blood plasma and urine

All biofluid samples were collected in the morning, after overnight fasting, in order to minimise dietary influence, although no control over previous food intake was performed.

Blood was collected by venipuncture into 9 mL sodium heparin tubes (Vacuette®) and centrifuged (1500 *g*, 10 min) within a maximum of 30 min to separate the plasma in order to avoid cell rupture/leakage and minimize transport of metabolites between intra and extracellular compartments (Teahan et al. 2006). Plasma aliquots of approximately 1 mL were then transferred into sterile cryovials, frozen and stored at -80 °C until analysis.

Urine samples were collected into sterile cups, centrifuged (1500 *g*, 10 min), divided into 1 mL aliquots, frozen and stored at -80ºC until analysis. The maximum storage period was four months for blood plasma and six months for urine.

2.3 NMR spectroscopy

2.3.1 Sample preparation for NMR

The experimental procedures described below were based on standardised protocols reported in the literature (Beckonert et al. 2007; Beckonert et al. 2010).

2.3.1.1 Lung tissue

At the time of analysis, each frozen tissue fragment was washed with a few drops of D_2O saline (0.9% NaCl) to remove excess blood (containing paramagnetic species that could lead to signal broadening), as well as water (for improved water suppression). About 40 mg of thawed cold tissue (38.7±7.2 mg) were then cut and packed into a 4 mm zirconium HRMAS rotor, fitted with a Teflon® top insert to provide a 50 μL sample volume with a cylindrical geometry. Ten microliters of D_2O saline with 0.25% of 3-(trimethylsilyl) propionate sodium salt (TSP)-*d*4 were also added to provide a signal for lock (D_2O) and for shimming (TSP).

2.3.1.2 Blood plasma

Plasma samples were thawed at room temperature, vortexed and their pH measured (7.81±0.13). Then, 400 μL of saline solution (NaCl 0.9% in 10% D_2O) were added to 200 μL of plasma, and the pH measured again (7.85±0.14). After centrifugation (4293 *g*, 5 min), 550 μL of the supernatant were transferred into a 5 mm NMR tube.

2.3.1.3 Urine

At the time of NMR analysis, urine samples were thawed at room temperature and their pH measured (6.07±0.86). Samples were then centrifuged (4293 *g*, 5 min) to remove residual cells and other precipitated material. Afterwards, 540 μL of supernatant were mixed with 60 μL of buffer solution (KH_2PO_4 1.5 M in D_2O) containing 0.1% TSP-*d*4 (used as chemical shift reference) and 2 mM of sodium azide (NaN_3, bacteriostatic agent). The pH of the buffered sample was again measured (6.78±0.15), and if necessary a final adjustment to pH 7.01±0.01 was carried out by adding a few drops of 4 M KOD or DCl. Finally, samples were centrifuged (4293 *g*, 5 min) and 550 μL of supernatant were transferred into a 5 mm NMR tube.

2.3.2 Acquisition and processing of HRMAS NMR spectra of tissues

NMR spectra of intact tissue samples were acquired on a Bruker Avance DRX-500 spectrometer (University of Aveiro, Portuguese NMR network, Portugal) operating at 500.13 MHz for ^{1}H observation, using a 4 mm HRMAS probe (g-HR MAS 500 SB BL4). The rotor containing the sample was spun at the magic angle (54.7º relative to the magnetic field) at a 4 kHz spinning rate. A Bruker cooling unit was used to cool down the bearing air flowing into the probe in order to keep the temperature at 277 K.

Three ^{1}H 1D NMR experiments were acquired for each sample: i) standard ^{1}H NMR with water suppression by presaturation during the relaxation delay (RD) and mixing time (t_m) (*noesypr1d* pulse programme from Bruker library, RD-90º-t_1-90º-t_m-90º-ACQ, t_1 being a short delay of 3 μs), ii) T_2-edited (Carr-Purcell-Meiboom-Gill, CPMG) ^{1}H NMR for attenuation of broad signals from fast-relaxing molecules (*cpmgpr* pulse programme from Bruker library, RD-90º-(τ-180º-τ)$_n$-ACQ, τ being the echo time and n the number of loops), and iii) diffusion-edited ^{1}H NMR for selecting signals of bound or large slow-diffusing molecules (*ledbpgp2s1dpr* pulse programme from Bruker library, RD-90º-G1-180º-G1-

90º-G2-τ-90º-G1-180º-G1-180º-G1-90º-G2-τ-90º-ACQ, G1 and G2 being the pulsed-field and the spoil gradients, respectively, and τ the gradient recovery delay). The main acquisition parameters are listed in Table 2.2. Acquisition time for each FID was set to 2.52 s (to provide adequate digital resolution) and the relaxation delay to 4 s, giving a total recycle time of 6.52 s. Based on ^{1}H T_1 measurements (described in section 2.3.4 and shown in Annex VI) these conditions allowed semi-quantitative standard 1D spectra to be recorded within a reasonable measurement time (approximately 29 min per sample).

Spectral processing was performed using Topspin 3.2 (Bruker Biospin, Rheinstetten, Germany). Prior to Fourier transform (FT), the FIDs were zero-filled to 64 k points and multiplied by an exponential line-broadening window function (LB 0.3 Hz). All 1D spectra were manually phased and baseline corrected. Chemical shifts were referenced internally to the alanine signal at δ 1.48, since this peak was found to be more reliable than the TSP singlet (often broadened or shifted probably due to interaction with macromolecules). Main processing parameters of 1D spectra are also listed in Table 2.2.

Table 2.2 Main parameters used for the acquisition and processing of 500 MHz 1D ^{1}H HRMAS NMR spectra of intact lung tissue.

	^{1}H HRMAS NMR		
Acquisition parameters			
Experiment	standard 1D	CPMG	diffusion-edited
Pulse programme[a]	*noesypr1d*	*cpmgpr*	*ledbpgp2s1dpr*
Number of scans, NS	256	256	256
FID data points, TD	32768	32768	32768
Spectral width (ppm)	13.02	13.02	13.02
Acquisition time, ACQ (s)	2.52	2.52	2.52
Relaxation delay, RD (s)	4	4	4
Mixing time, t_m (ms)	100		
Spin echo delay, τ (μs)		300	
Number of loops, n		150	
Pulsed-field gradient, G1 (μs)			200
Diffusion delay, Δ (s)			0.1
Delay for longitudinal eddy currents, τ (ms)			5
Spoil-gradient, G2 (ms)			2
Gradient pulse, δ/2 (μs)			2000
Processing parameters			
Window function	exponential	exponential	exponential
Spectrum data points, SI	65536	65536	65536
Line broadening, LB (Hz)	0.3	0.3	0.5

[a] Bruker library (pulse programmes can be found in Annex VII).

Two-dimensional (2D) ^{1}H-^{1}H total correlation (TOCSY) spectra, ^{1}H-^{13}C phase sensitive heteronuclear single quantum correlation (HSQC) spectra and *J*-resolved spectra were registered for selected samples to aid spectral assignment. Bruker Avance DRX-500 (University of Aveiro, Portuguese NMR network, Portugal) and Bruker Avance II 600 (Bruker Biospin, NMR Division, Rheinstetten, Germany) spectrometers, operating respectively at 500.13 and 600.13 MHz for ^{1}H observation were used to record 2D spectra. Typically, TOCSY spectra were acquired in the phase-sensitive mode using time proportional phase incrementation (TPPI) and the MLEV 17 pulse sequence for the spin-lock (*mlevgsst* and *mlevphpr* pulse programmes from Bruker library), ^{1}H-^{13}C phase sensitive (echo/antiecho) HSQC were recorded with inverse detection and carbon decoupling (*hsqcetgppr* and *invietgpsi* pulse programmes from Bruker library) and *J*-resolved spectra with F2 decoupling (*jresgpprqf* pulse programme from Bruker library). For both TOCSY and HSQC spectra, zero filling to 1024 data points and forward linear prediction were used in F1 dimension and multiplication by a shifted sine-bell-squared apodization function was applied in both dimensions prior to FT and phasing. For the *J*-resolved experiments, prior to FT, the FIDs were weighted in both dimensions by a sine-bell function and zero-filled in the F1 dimension to 256 data points. The spectra were tilted by 45º to provide orthogonality of the chemical shift and coupling constant axes were subsequently symmetrised about the F1 axis. Main acquisition and processing parameters of 2D HRMAS NMR spectra acquired for lung tissues are summarised in Table 2.3.

Table 2.3 Main parameters used for the acquisition and processing of 2D HRMAS NMR (600/500 MHz) spectra of intact lung tissues.

	^{1}H HRMAS NMR (600 MHz/500 MHz)		
Acquisition parameters			
Experiment	^{1}H-^{1}H TOCSY	^{1}H-^{13}C HSQC	*J*-Resolved
Pulse programme[a]	*mlevphpr/mlevgsst*	*hsqcetgppr/invietgpsi*	*jresgpprqf*[b]
FID data points 1st dimension [F1]	2048/1024	1024/1024	16384
FID data points 2nd dimension [F2]	280/256	200/256	40
Number of scans	80/32	88/60	48
Spectral width [F1] (ppm)	12.02/13.02	20.55/13.02	16.66
Spectral width [F2] (ppm)	12.02/123.02	190.00/249.99	0.13
Relaxation delay, RD (s)	2/2	4/2	2
Mixing time, t_m (ms)	30/80		
Processing parameters			
Spectrum data points [F1], SI	2048/1024	1024/1024	16384
Spectrum data points [F2], SI	1024/1024	1024/512	256

Table 2.3 (continued)

		^{1}H HRMAS NMR (600 MHz/500 MHz)	
Processing parameters			
Line broadening [F1], LB (Hz)	0.3/0.3	0.3/0.3	0.3
Line broadening [F2], LB (Hz)		1.0/1.0	0.3

[a] Bruker library; [b] spectra were acquired in the 600 MHz spectrometer only.

2.3.3 Acquisition and processing of NMR spectra of biofluids

NMR spectra of urine and blood plasma were acquired on a Bruker Avance DRX-500 spectrometer (University of Aveiro, Portuguese NMR network, Portugal) operating at 500.13 MHz for ^{1}H observation, at 300 K, using a 5 mm BBI probe (BBI 500 MHz SB 5 mm).

In the case of plasma three 1D ^{1}H NMR experiments were acquired for each sample, similarly to what has been previously described for tissues: i) standard 1D with water suppression (*noesypr1d* pulse programme from Bruker library), ii) T_2-edited (Carr-Purcell-Meiboom-Gill, CPMG) (*cpmgpr* pulse programme from Bruker library) and iii) diffusion-edited (*ledbpgp2s1dpr* pulse programme from Bruker library). In the case of urine, only the standard 1D ^{1}H NMR spectrum was recorded for each sample, given that urine is typically free of macromolecular components, thus sparing the need for relaxation- or diffusion-edited experiments to be recorded. All 1D spectra were manually phased and baseline corrected. The chemical shifts were referenced internally either to the α-glucose signal at δ 5.23 in the case of plasma and to the TSP signal at δ 0.00 in the case of urine. The main acquisition and processing parameters can be found in Table 2.4.

Table 2.4 Main parameters used for the acquisition and processing of 500 MHz 1D ^{1}H NMR spectra of blood plasma and urine.

	Blood plasma			**Urine**
Acquisition parameters				
Experiment	standard 1D	CPMG	diffusion-edited	standard 1D
Pulse programme[a]	*noesypr1d*	*cpmgpr*	*ledbpgp2s1dpr*	*noesypr1d*
FID data points, TD	32768	32768	32768	32768
Number of scans	128	256	128	128
Spectral width (ppm)	20.66	20.66	20.66	20.66
Relaxation delay, RD (s)	4	4	4	4
Mixing time, t_m (ms)	100			100
Spin echo delay, τ (μs)		400		

Table 2.4 (continued)

	Blood plasma			Urine
Processing parameters				
Number of loops, n		80		
Pulsed-field gradient, G1 (μs)			100	
Diffusion delay, Δ (s)			0.1	
Delay for longitudinal eddy currents, τ (ms)			5	
Spoil-gradient, G2 (ms)			2	
Gradient pulse, δ/2 (μs)			1000	
Window function	exponential	exponential	exponential	exponential
Spectrum data points, SI	65536	65536	65536	65536
Line broadening, LB (Hz)	0.3	0.3	0.5	0.3

[a] Bruker library (pulse programmes can be found in Annex VII).

2D homonuclear (^{1}H-^{1}H TOCSY and *J*-resolved) and heteronuclear (^{1}H-^{13}C HSQC) spectra were also registered for selected biofluid samples to aid spectral assignment. Main acquisition and processing parameters of 2D experiments are listed in Table 2.5.

Table 2.5 Main parameters used for the acquisition and processing of 500 MHz 2D NMR spectra of blood plasma and urine.

	Blood plasma/Urine		
Acquisition parameters			
Experiment	^{1}H-^{1}H TOCSY	^{1}H-^{13}C HSQC	J-resolved
Pulse programme[a]	*clmlevprtp*	*invietgpsi*	*lcjresprqf*
FID data points [F1]	240/256	360/400	80/80
FID data points [F2]	2048/2048	3072/4096	3072/3072
Number of scans	48/56	40/40	32/32
Spectral width [F1] (ppm)	20.66/16.00	200/200	0.19/0.19
Spectral width [F2] (ppm)	20.66/16.02	20.66/16.02	20.66/20.66
Relaxation delay, RD (s)	1.5	1.5	2/1
Mixing time, t_m (ms)	80		
Processing parameters			
Window function	qsine/qsine	qsine/qsine	sine
Spectrum data points [F1], SI	1024/1024	1024/512	16384/16384
Spectrum data points [F2], SI	4096/4096	4096/4096	4096/4096
Line broadening [F1], LB (Hz)	0.3/0.3	0.3/0.3	0.3
Line broadening [F2], LB (Hz)		1.0/1.0	0.3

[a] Bruker library.

2.3.4 ^{1}H T_1 and T_2 measurements

Proton relaxation parameters, namely spin-lattice (longitudinal) relaxation time constant, T_1, and spin-spin (transverse) relaxation time constant, T_2, were measured for selected samples of tissues and biofluids. For measuring ^{1}H T_1 values a standard inversion-recovery pulse sequence (RD-180°$_x$-τ_1-90°$_x$-ACQ) was used, while ^{1}H T_2 values were measured using a spin echo experiment (RD-90°-(τ-180°-τ)$_n$-ACQ). The main acquisition and processing parameters are listed in Table 2.6.

Table 2.6 Main parameters used for the acquisition and processing of ^{1}H T_1 and T_2 measurements.

	Lung tissue	**Blood plasma**	**Urine**
Acquisition parameters			
Experiment		*T_1 measurement*	
Pulse programme[a]	*t1irpr*	*t1irpr*	*t1irpr*
FID data points [F1]	26	32	32
FID data points [F2]	16384	16384	16384
Number of scans	40	64	64
Spectral width [F1] (ppm)	2.00	2.00	2.00
Spectral width [F2] (ppm)	13.02	19.99	19.99
Relaxation delay, RD (s)	10	10	10
Range of delays (ms)	0.01-20.00	0.01-25.00	0.01-25.00
Experiment		*T_2 measurement*	
Pulse programme[a]	*t2zgpr*	*t2zgpr*	
FID data points [F1]	30	32	
FID data points [F2]	16384	16384	
Number of scans	40	64	
Spectral width [F1] (ppm)	2.00	2.00	
Spectral width [F2] (ppm)	13.02	19.99	
Relaxation delay, RD (s)	10	10	
Range of loops	5-8000	2-15000	
Processing parameters (T_1/T_2)			
Window function	sine/sine	sine/sine	Sine/
Spectrum data points [F1], SI	16384/16384	16384/16384	32768/
Spectrum data points [F2], SI	26/30	32/32	32/
Line broadening [F1], LB (Hz)	0.3	0.3	0.3

[a] Bruker library (pulse programmes can be found in Annex VII).

^{1}H T_1 and T_2 values were estimated by fitting experimental data using OriginPro 8 (OriginLab, Northampton, Massachusetts, USA) to exponential decay functions shown in

Equation 1.10 and Equation 1.11, respectively. For some spectral regions with overlap of sharp and broad signals, a single-exponential decay did not fit the data, thus bi-exponential decay functions were applied.

2.3.5 Pre-treatment and multivariate analysis of NMR spectra

In order to prepare the data for multivariate analysis (MVA), full resolution ^{1}H NMR spectra (no bucketing) were organised as rectangular $n \times m$ (rows×columns) matrices of n observations (samples) and m variables (peak intensities), designated X matrices (main features shown in Table 2.7), by using Amix 3.9.14 (Bruker Biospin, Rheinstetten, Germany). Selected regions were excluded from these matrices, namely the suppressed water signal (in all sample types), ethanol and glycol contaminant signals in tissue spectra (probably arising, respectively, from the desinfected surgical instruments and the cryopreservative embedding medium), and the urea signal in urine spectra (close to water and affected by water suppression via proton exchanging).

The rows in those were then aligned using a recursive segment-wise peak alignment (RSPA) algorithm (Veselkov et al. 2009) and normalized by probabilistic quotient normalisation (PQN), using the median of the controls as reference spectrum (Dieterle et al. 2006). Both procedures were carried out in MATLAB version 7.14.0.739 (The MathWorks, Inc., Natick, Massachusetts, USA), using scripts developed at ICL and kindly supplied by Dr. Kirill A. Veselkov. Afterwards, the columns were mean centered (subtraction by average value) and different scaling types were tested (none – just centered, unit variance and pareto scaling), using SIMCA-P 11.5 software (Umetrics, Umeå, Sweden). UV scaling was found to give the best results for most models built.

Multivariate analysis was carried out using SIMCA-P version 11.5 (Umetrics, Umeå, Sweden) and comprised principal component analysis (PCA), partial-least squares-discriminant analysis (PLS-DA) and orthogonal projection to latent structures-discriminant analysis (OPLS-DA). The results were visualized through scores scatter plots and corresponding loadings plots. In the case of models involving UV scaling, the loadings profiles were recovered by multiplying the loading weights w by the standard deviation. Colouring by the variable importance in the projection (VIP) was then performed in R software version 2.15.0 (R Development Core Team, Vienna, Austria, 2012). In the case of models with low predictive power, a variable selection method developed in-house was

applied (Diaz et al. 2013). According to this method, the variables kept for further MVA and model validation simultaneously obeyed the following criteria: i) VIP>1; ii) VIP/ VIP_{cvSE}>1 and iii) $|b/b_{cvSE}|$>1, where VIP_{cvSE} and b_{cvSE} are, respectively, the standard errors of VIP values and *b* coefficients, obtained in PLS-DA models.

Table 2.7 Information about the ^{1}H NMR matrices used for MVA.

Sample type	Lung tissue	Blood plasma	Urine
Rows, n samples[a]	115	200	200
Columns, m variables (intensities)			
Standard 1D	39539	35961	25125
CPMG	37776	22476	n.a.
Diffusion-edited	20590	36325	n.a.
Spectral interval (ppm)			
Standard 1D	0.25-8.80	-1.50-10.50	0.50-9.40
CPMG	0.50-8.80	0.50-8.50	n.a.
Diffusion-edited	0.25-4.38	-1.50-10.50	n.a.
Exclusion areas (ppm)	ethanol (1.15-1.22, 3.63-3.69)[b] glycol (3.69-3.73) water (4.85-5.20)	water (4.60-4.85)	water (4.62-4.90) urea (5.50-6.10)
Integration mode	Sum of absolute intensities		

n.a. Not applicable. [a] Number of samples in controls vs. cancer MVA models. [b] Not detected in diffusion-edited spectra.

2.3.6 Model validation

The robustness of PLS-DA classification models was assessed by Monte Carlo cross validation (MCCV) and permutation testing (500 iterations) as explained in subchapter 1.3.4.4, using an in-house software developed and kindly supplied by Dr. António Barros. For each model, a ROC space plot of the true positive rate (TPR or sensitivity) against the false positive rate (FPR or 1-specificity), together with a Q^2 histogram showing the distribution of Q^2 values in original (true classes assigned) and permuted model iterations, were used to assess robustness. In particular, PLS-DA models were considered validated when there was no overlap in the ROC plot between models obtained for true and permuted classes, the true models showing high sensitivity and specificity, and there was minimal overlap between original and permuted Q^2 values, the median Q^2 being high (>0.5) for original models and low or negative for permuted models.

In the case of biofluids, preliminary external validation was carried out to evaluate the performance of MVA models for predicting the class of new samples (SIMCA-P version 11.5, Umetrics, Umeå, Sweden). The data was divided into training (or calibration)

and test (or validation) sets, the later consisting of forty samples most recently collected and analysed (and not included in the training set). Prediction was performed by using either the whole dataset (all variables) or the signals found to be most relevant for the discrimination between controls and cancer patients.

2.3.7 Spectral integration and univariate statistics

To evaluate individual metabolite quantitative variations, selected signals (with VIP>1) were integrated in the ^{1}H NMR spectra and normalised by the corresponding PQN quotient. Spectral integration was performed by calculating the area under a certain signal for which integration limits were manually defined or, in cases of highly overlapped regions, by applying deconvolution using Amix-Viewer 3.9.14 software (Bruker BioSpin, Rheinstetten, Germany). Deconvolution using an automated variable Gaussian/Lorentzian peak fitting routine based on the Marquardt algorithm was applied to the region of interest, and the resulting peak fits were improved by manually adjusting the peak picking until the error estimates associated to the intensity, the height at half-width, and the Gaussian/Lorentzian proportion were minimised.

Spectral integrals were statistically compared by applying the Wilcoxon signed rank test for paired samples (lung tissues) and the Wilcoxon rank sum test for unpaired samples (biofluids). Additionally, Bonferroni correction was applied to minimise false discoveries (false positives) in multiple comparisons by adjusting the significance level as a function of the numbers of variables tested. Also, for each metabolite, the percentage of variation and respective error were calculated, along with the effect size (*d*) and respective 95% confidence interval (according to Equation 1.30, Equation 1.31 and Equation 1.32).

2.4 Ultra-performance liquid chromatography – mass spectrometry (UPLC-MS)

2.4.1 Sample preparation for UPLC-MS

For UPLC-MS analysis, urine samples were thawed at room temperature and centrifuged (9500 *g*, 10 min) before dilution of the supernatant (1:1 v/v) with HPLC grade water. All samples were vortexed, transferred to 96-well plates and kept in the UPLC auto-sampler unit at a constant temperature of 4º C throughout the entire run. Quality control

(QC) samples were also prepared by mixing 10 μL aliquots of all urine samples used in this study.

2.4.2 Acquisition and processing of UPLC-MS data

UPLC-MS analysis of urine was performed in the Biomolecular Medicine Department, Division of Surgery and Cancer of the Faculty of Medicine, Imperial College of London (ICL), United Kingdom. An Acquity UPLC® System (Waters Corporation, Milford, Massachusetts USA) coupled to a LCT Premier ToF mass spectrometer (Waters Corporation, Milford, Massachusetts USA) was used.

With the purpose of achieving an extensive coverage of the urinary metabolic content, two complementary chromatographic columns were used in two separate runs: a reverse-phase high-strength silica column (2.1×100 mm, 1.8 μm, HSS T3 Acquity UPLC®, Waters Corporation, Milford, Massachusetts USA) and a hydrophilic interaction chromatography column (2.1×100 mm, 1.7 μm, HILIC BEH Acquity UPLC®, Waters Corporation, Milford, Massachusetts USA). Compared to traditional C18 columns, the HSS T3 column shows improved performance regarding retention of certain polar metabolites (New and Chan 2008), while the HILIC BEH column has the ability of improving the profiling of highly polar metabolites (Gika et al. 2008; Spagou et al. 2011). Mass spectrometry detection was separately performed in both positive and negative ionization modes. Therefore, each sample was analysed four times, using either positive or negative ionization mode in each column. The mass analyser was a time-of-flight (ToF) analyser interfaced with an electrospray ion source.

General experimental conditions of liquid chromatography were applied as described by Want et al. (Want et al. 2010) and are summarised in Table 2.8. Column temperature was maintained at 40ºC and eluent flow rate kept at 0.5 $mL \cdot min^{-1}$ throughout the analysis. The sample volume injected was 5 μL and a different elution gradient was applied to the two columns (Table 2.8). For the HSS column the eluent was composed of water with 0.1% formic acid (A_{HSS}), and acetonitrile with 0.1% formic acid (B_{HSS}), while for the HILIC column, the eluent consisted of acetonitrile with 5% ammonium acetate (A_{HILIC}) and acetonitrile with 50% ammonium acetate (B_{HILIC}).

Table 2.8. Experimental conditions of UPLC urine analysis.

Column type	**HSS**		**HILIC**	
Main parameters				
Column temperature (ºC)	40		40	
Injection volume (μL)	5		5	
Flow rate ($mL \cdot min^{-1}$)	0.5		0.5	
Mobile phase (%)	A: 0.1% HCOOH in water B: 0.1% HCOOH in acetonitrile		A: 5% CH_3COONH_4 in acetonitrile B: 50% CH_3COONH_4 in acetonitrile	
Time of run (min)	12		15	
Elution gradient				
Time (min)	A_{HSS} (%)	B_{HSS} (%)	A_{HILIC} (%)	B_{HILIC} (%)
0.0	100	0	99	1
1.0	100	0	99	1
3.0	85	15	99	1
6.0	50	50	99	1
9.0	5	95	99	1
10.0	5	95	99	1
10.1	100	0	99	1
12.0	100	0	0	100
12.1			99	1
15.0			99	1

Column conditioning was performed by running, at the beginning of each analysis, ten pooled QC samples, prepared as described in section 2.4.1, in order to achieve stable retention times (RT) of the compounds. Moreover, QC samples were also injected every ten samples throughout the analytical run to allow stability and quality of the data to be assessed (Want et al. 2010). For further evaluation of data quality (RT, peak shape, signal intensity) a test mixture, comprising twenty five reference compounds, and a blank (water, also used for the mobile phases) were run at the beginning and at the end of the analysis. Urine samples of control and cancer groups were analysed in random order.

Mass spectrometry was performed in both positive and negative ion electrospray modes (ESI+ and ESI-), and the general ionisation parameters are listed in Table 2.9. The capillary voltage was 3.0 kV (ESI+) and 2.5 kV (ESI-), and the cone voltages were set to 30 V (ESI+) and 25 V (ESI-). The eluent desolvation temperature was 400 ºC, while the source temperature was 120 ºC. Finally, the cone gas flow rate was 25 $L \cdot min^{-1}$ and the desolvation gas flow rate was 800 $L \cdot h^{-1}$. The LCT Premier was operated in V optics mode with a scan time of 0.1 s and an interscan delay of 0.01 s. For exact mass measurements, a LockSpray interface was used to correct changes in environment or experimental

conditions over the course of the analysis. The reference, or lock mass, used was a 200 $\mu g \cdot L^{-1}$ leucine enkephalin (555.2645 amu) peptide solution (50/50 acetonitrile/water with 0.1% v/v formic acid) sprayed at a flow rate of 3 $\mu L \cdot min^{-1}$. Data were collected in centroid mode with a scan range of 50-1000 *m/z*, with lock mass scans being collected every 15 scans and averaged over 5 scans to perform mass correction.

Table 2.9. MS ionisation parameters for ESI+ and ESI- modes.

Ionisation mode	**ESI+**	**ESI-**
Voltages (V)		
Capillary	3000	2500
Sample cone	30	25
Temperatures (ºC)		
Desolvation	400	
Source	120	
Gas flows		
Cone ($L \cdot min^{-1}$)	25	
Desolvation ($L \cdot h^{-1}$)	800	

2.4.3 Pre-treatment and multivariate/univariate analysis of UPLC-MS data

UPLC-MS data pre-treatment was carried out in R software version 2.15.0 (R Development Core Team, Vienna, Austria, 2012), using the freely available software package XCMS (Smith et al. 2006), and consisted of several steps as shown in Figure 2.1.

Raw data was converted to netCDF format using Waters MassLynx™ version 4.1 Databridge software (Waters Corporation, Milford, Massachusetts USA) and then imported to R software for processing. Scripts used here were developed at ICL and kindly supplied by courtesy of Dr. Matthew Lewis and Dr. Paul Benton. To carry out peak detection, the *centWave* peak picking algorithm for high resolution datasets was used (Tautenhahn et al. 2008), together with a pre-filter to remove mass traces that were not detected in at least 5 scans with intensity higher than a 1000. This filter allowed for some of the background noise to be reduced. The peakwidth (minimum-maximum) used for peak picking was 2-15 s for HSS and 3-60 s for HILIC columns.

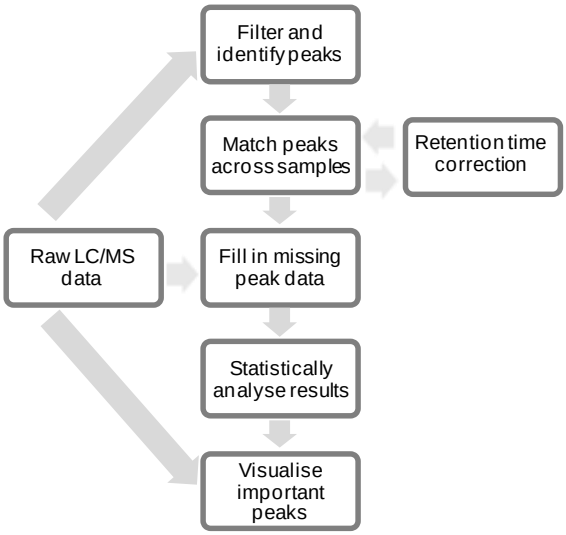


Figure 2.1. Schematic workflow for pre-treatment of LC-MS data (adapted from Smith et al. 2006)**.**

After peak detection in individual samples, peaks must be matched across samples to allow the calculation of retention time deviations. Hence, peaks within a *m/z* bandwidth of 0.05 Da were firstly grouped within an interval of 5 s relatively to the median RT of 'well-behaved' peaks, and retention times' deviations were calculated. Subsequently, a second, more accurate, peak alignment was performed taking into account the new RT deviations obtained, namely 4 and 3 s for HSS positive and negative modes, and 6 and 12 s for HILIC positive and negative modes, respectively. Missing peaks were identified and their intensity measured in raw data.

A following MinFrac filter was used, in which features not present in at least 50% of a sample group (QCs, control and cancer groups) were filtered out and considered to be noise. This filter was found to be very effective in removing background features, as a previous attempt to process data without applying MinFrac filter resulted in a final matrix with several background features. Normalisation of features' intensity was performed by applying the PQN method (Veselkov et al. 2011). At last, a coefficient of variation (CV)-based filter was applied, in which spectral features with a CV (%) higher than 30% were removed. The CV was calculated only for the QCs analysed between samples (not the QCs used for column conditioning). This filter guaranteed that only features with high reproducibility throughout the entire run were kept.

The normalized UPLC data sets were then arranged into $n \times m$ matrices, where n are the samples (rows) and m are the data features' intensities described by specific retention

time-*m/z* pairs. Table 2.10 shows the decreasing number of features present in the UPLC matrices with the application of the previously described filters. Datasets were then mean-centered and subjected to different scaling types (UV scaling, Pareto scaling or logarithmic transformation) before MVA.

Table 2.10 Information about the UPLC-MS matrices used for MVA.

	HSS ESI+	**HSS ESI-**	**HILIC ESI+**	**HILIC ESI-**
Rows, n samples				
QC		11		
Control		50		
Cancer		48		
Columns, m variables (RT_m/z pairs)				
Raw data	19323	36620	11260	12853
After MinFrac filter	4959	8226	3157	3436
After CV filter	4434	7312	2628	2412

MVA was carried out in SIMCA-P version 11.5 (Umetrics, Umeå, Sweden). PCA and PLS-DA were applied to UPLC-MS data, and extraction of relevant features from validated PLS-DA models was performed by applying consecutive selection rules to the loadings list. Firstly, S-plots relating covariance (p) and correlation (p_{corr}) between the UPLC-MS variables and the PLS-DA scores were analysed so that only variables with $p>0.02$ and $|p_{corr}|>0.6$ were considered relevant. Secondly, the ratio of the absolute mean covariance loading and its respective jack-knifed cross-validated standard error ($p/p_{CV}SE$) was calculated and a threshold of 4 ($p_{CV}SE$ below 25%) for keeping features was established (Wiklund et al. 2008). Finally, a significance test was applied as explained before (subchapter 1.3.4.5).

As for NMR data, model validation was then performed by MCCV and permutation testing to assess PLS-DA models' robustness and predictive ability.

2.5 Statistical correlation: STOCSY and SHY

Statistical total correlation spectroscopy (STOCSY) analysis was applied to NMR spectra for helping the assignment of unknown resonances. Pearson correlation coefficients (r) were calculated between each unknown signal (driver peak) and the full spectral matrix and used to colour the covariance plot, resulting in a 1D STOCSY plot (Cloarec et al. 2005). To aid the identification of RT_*m/z* features, statistical heterospectroscopy (SHY)

was applied (Crockford et al. 2006). In this case, the correlation between each MS (RT_*m/z*) feature and the full NMR matrix (same samples) was calculated. Statistical significance of Pearson correlation coefficients was calculated and a threshold of $|r|>0.8$ ($p<0.01$) was applied in agreement with previous reports using STOCSY (Crockford et al. 2006; Maher et al. 2011).

3 METABOLIC PROFILING OF LUNG TUMOURS BY TISSUE NMR METABOLOMICS

This chapter presents the analysis of tumour and non-involved (control) lung tissues by ^{1}H HRMAS NMR spectroscopy. In the first subchapter, some experimental details are discussed, namely regarding the influence of spinning rate, time and temperature on the spectral profile of lung tissues, and the use of relaxation- and diffusion-edited experiments to deal with spectral overlap. Subsequently, detailed metabolite assignment is pursued with the help of 2D NMR spectroscopy and STOCSY. In the third and fourth subchapters, MVA is applied to search for metabolic features differentiating lung tumour and control tissues, as well as tumour histological types. Finally, the biochemical interpretation of the metabolic variations observed is proposed.

3.1 ^{1}H HRMAS NMR spectra of lung tissues: setting up the acquisition conditions

The typical standard 1D ^{1}H HRMAS NMR spectrum of lung (tumour) tissue, with water signal suppression, is shown in Figure 3.1. This spectrum comprises well resolved peaks across the whole spectral range (line widths at half-height $\leq$ 2 Hz for most signals) overlapped with some broad resonances arising from lipid and protein components. Before going into detail about the assignment of spectral peaks to specific compounds, the following sections will address the variations observed in lung tissue NMR profile upon sample spinning at different rates, temperatures and time intervals. Moreover, the optimization and use of relaxation- and diffusion-edited experiments to deal with the spectral overlap observed in the standard 1D spectrum will be described. These preliminary tests were important to establish the experimental conditions for acquiring informative, reproducible and comparable sample spectra, especially given that no reports on the HRMAS analysis of intact lung tissue were available at the beginning of this work.

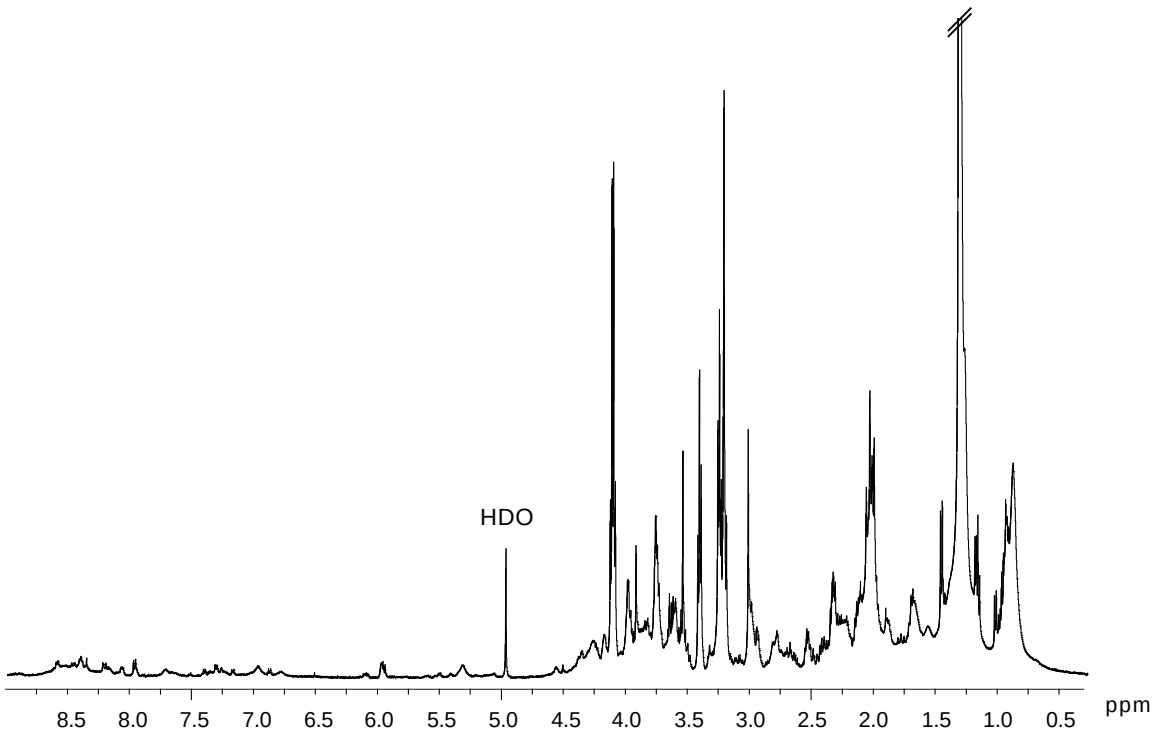


Figure 3.1 600 MHz standard 1D ^{1}H HRMAS NMR spectrum of lung tumour tissue.

3.1.1 Influence of spinning rate on the NMR spectral profiles

As described in section 1.3.2.4, spinning the rotor containing the tissue sample at the magic angle (54.74º in relation to the external magnetic field B_0) and at a few kHz spinning rate, allows line broadening effects to be minimized and well resolved spectra to be obtained from intact tissues.

In order to define the most suitable spinning rate (SR) for recording intact lung tissue spectra, consecutive spectra of the same sample were acquired at different SRs, varying from 1 to 8 kHz, as shown in Figure 3.2. Probe temperature was kept at 283 K, as this was the minimum value achievable at the lowest spinning rate. In general, a significant peak narrowing and improvement in resolution was observed when increasing SR from 1 to 2 kHz, while for the other spinning rates tested the line widths of signals arising from small metabolites remained fairly constant and showed little enhancement in spectral resolution (Table 3.1). On the other hand, broad resonances from lipids (e.g., at δ 0.89 and δ 1.26), which were poorly visible at 1 and 2 kHz spinning rates, became more resolved with increasing SR, up to 8 kHz (maximum SR tested, Figure 3.2). This observation suggests

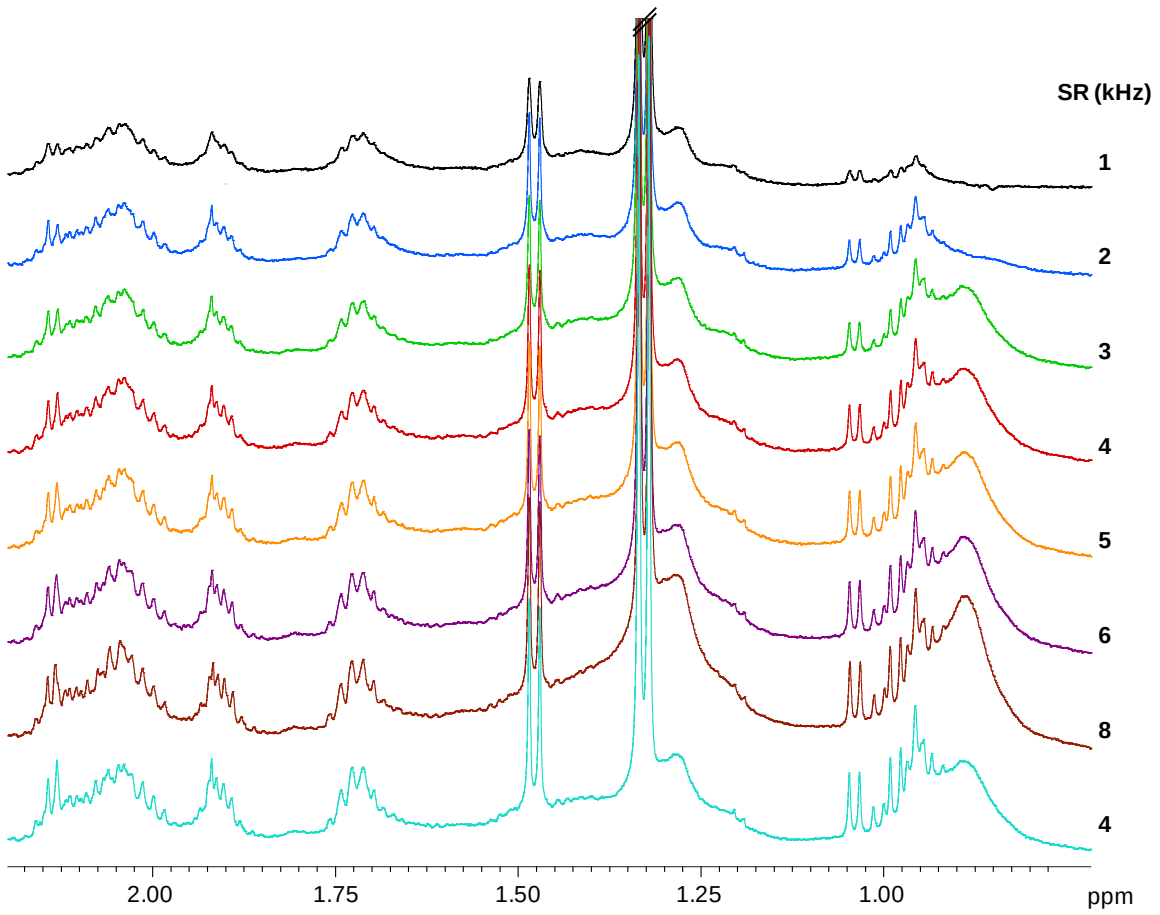


Figure 3.2 Expansion (δ 2.25-0.65) of consecutive standard 1D ^{1}H HRMAS NMR spectra of lung tissue acquired with different spinning rates.

Table 3.1 Widths at half-height (ν½) measured for several peaks at different spinning rates (shown in order of spectra acquisition).

	$\nu_{½}$ [a] (Hz)					
Spinning rate (kHz)	valine (δ 1.05)	lactate (δ 1.33)	alanine (δ 1.48)	*s*-inositol (δ 3.35)	taurine (δ 3.43)	glycine (δ 3.56)
1	3.44	3.24	3.37	3.10	4.24	3.26
2	2.27	2.10	2.02	2.10	2.78	1.69
3	2.22	2.23	2.23	2.26	2.83	1.88
4	2.12	2.14	2.13	2.13	2.73	1.83
5	2.07	2.01	2.07	2.10	2.64	1.75
6	2.14	2.10	2.18	2.21	2.78	1.88
8	2.12	2.01	2.23	2.07	2.69	1.86
4	2.01	1.88	1.91	1.99	2.49	1.64

[a] Values measured in spectra processed with LB 0 Hz.

that lipid resonances are affected by line broadening interactions (e.g., residual dipolar couplings) of higher magnitude than smaller molecules, possibly in relation to their more restrained molecular mobility, as indicated by their lower transverse relaxation time constants (T_2), compared to those of small molecules (Annex VI, Table A8).

In addition to line narrowing, a gradual increase in intensity was noted for all signals with increasing spinning rate, probably reflecting the progressive disruption of tissue integrity. Indeed, while in native tissue structure some metabolites may be partially bound to macromolecules or confined to compartments of low molecular mobility, consequently affording poor NMR visibility, spinning at high rates may facilitate their release into more fluid environments, thereby increasing their mobility and signal intensities. This effect is corroborated by the fact that signal intensities remained high when the spinning rate was decreased back to 4 kHz (Figure 3.2). A similar observation has been previously noted for other tissue types (Martinez-Granados et al. 2006; Swanson et al. 2008; Opstad et al. 2008). On the other hand, lipid resonances showed higher intensity in the spectrum recorded at 8 kHz than in the subsequent spectrum at 4 kHz (Figure 3.2), which is consistent with the anisotropic effects affecting their visibility, as previously explained.

Another variation observed in the spectra recorded at different spinning rates regards the small shift to lower frequencies in the water resonance (ca. 0.07 ppm) and in some metabolite signals, namely lactate, choline-containing metabolites, creatine and taurine (0.001-0.003 ppm). This shift is likely to reflect the increase in temperature arising from frictional heating with increasing SR, as temperature influences the strength of hydrogen bonds, thus proton shielding and chemical shifts. Indeed, by using Equation 3.1 (Garrod et al. 1999), the sample temperature was found to vary in the range 290-291 K for SR below 5 kHz, and to increase to 293 and 296 K for SR 6 and 8 kHz, respectively.

$$T = 256.87 + 84.17\Delta\alpha + 17.23\Delta\alpha^2$$ **Equation 3.1**

where $\Delta\alpha = \delta(H_1, glucose\ \alpha) - \delta(H_2O)$.

Based on the results presented above, 4 kHz was the spinning rate chosen for HRMAS analysis of all tissue samples, as it represented a good compromise between reduction of anisotropic interactions, preservation of tissue integrity and also exclusion of spinning side bands from the spectral region of interest, as for low SR (1-3 kHz) these

were located below 9 ppm, adding an extra difficulty to the analysis of an already complex, highly overlapped spectrum.

3.1.2 Stability of NMR spectral profiles during acquisition at different temperatures

The temperature used for storage and spectral acquisition of tissue samples is of paramount importance, as the excised tissue is no longer under the enzymatic control of the body, and both enzymatic and chemical degradation of metabolites may occur. So, after excision, lung tissue samples were immediately snap-frozen in liquid nitrogen and kept at -80 ºC until analysis. Samples used in this study were only thawed once just before NMR experiments, so no freeze-thaw effects were expected. Nonetheless, chemical and enzymatic reactions in tissue may become active once the sample is thawed and prepared for NMR analysis. Therefore, it is important to assess the extension of the metabolic variations during data acquisition, in order to ensure that the metabolic profiles are not strongly affected by uncontrolled degradation effects. In this work, great care was taken to keep the time between sample preparation and spectral recording as short and similar as possible for all samples (about 5 min at room temperature for rotor assembling and 10-15 min below 10 ºC for temperature stabilisation and optimization of NMR acquisition). Moreover, as recommended in the literature for other tissue types (Sitter et al. 2009), spectra were recorded at 277 K (4 ºC). As shown by the results presented below, this allowed the alterations in metabolite levels during acquisition time to be significantly minimized.

Tissue sample stability in the spectrometer was assessed by acquiring consecutive spectra during 5 hours for two fragments of the same lung tissue sample, one analysed at 277 K and the other at 293 K (SR 4 kHz in both cases). Then, PCA was applied to these spectra in order to highlight the main sources of variability. The resulting PC1 vs. PC2 scores scatter plot (Figure 3.3a) shows the separation between the two datasets (acquired at 277 and 293 K) in PC1 (mainly due to the temperature-related drifts in some peaks, as discussed in the previous section), while the distribution along PC2 reflects spectral changes with time. The dispersion of spectra along PC2 was much higher for the 293 K set than for the 277 K set, which is an indication of the larger variation in spectral profiles (i.e., reduced sample stability) at higher temperatures, as expected.

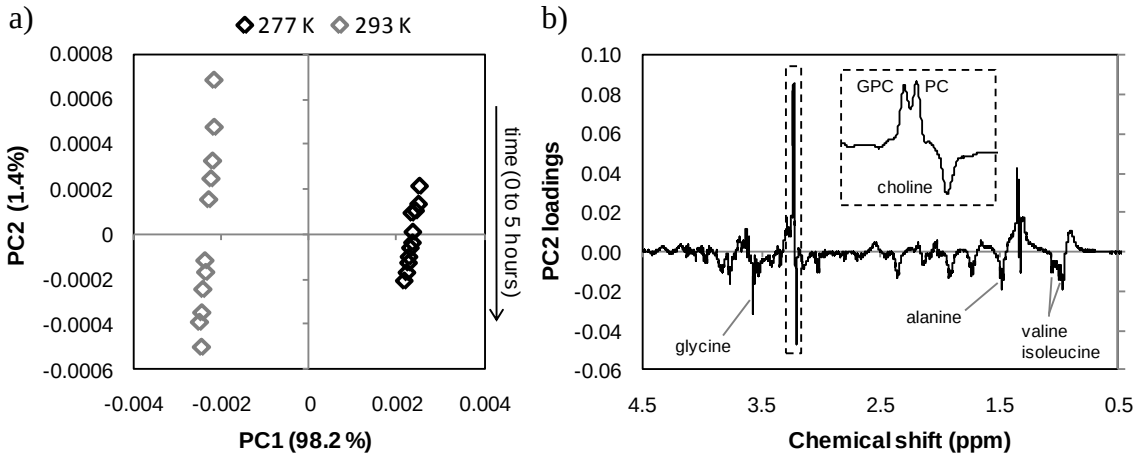


Figure 3.3 PCA of consecutive standard 1D ^{1}H HRMAS spectra (δ 0.5-4.5) of two fractions of the same lung sample recorded at 277 K (black) and 293 K (grey): a) PC1 vs. PC2 scores scatter plot and b) PC2 loadings plot. (GPC, glycerophosphocholine; PC, phosphocholine).

Inspection of PC2 loadings (Figure 3.3b) showed that phosphocholine and glycerophosphocholine decreased with time, whereas choline and several amino acids (alanine, valine, glycine, isoleucine) increased. Spectral integration was then employed to verify the magnitude of these variations, as shown in Figure 3.4. Notably, at 293 K, both GPC and PC decreased steadily by 10-11% over five hours, while choline increased 15%. These variations suggest PC and GPC degradation to choline, as already observed in prostate and brain tissues (Swanson et al. 2006; Swanson et al. 2008; Opstad et al. 2008). Moreover, several amino acids (glycine, alanine, valine, and isoleucine) exhibited a variation of 25-37% at 293 K after 5 h, possibly reflecting proteolytic activity in the tissue. Other variations such as those of lipids, lactate, acetate, *myo*- and *scyllo*-inositol were found to be residual (<5%) during the time window considered. At 277 K, on the other hand, the changes over five hours in choline-containing compounds were lower than 5%, and the only noticeable changes were seen for amino acid levels (15-23% increase) (Figure 3.4).

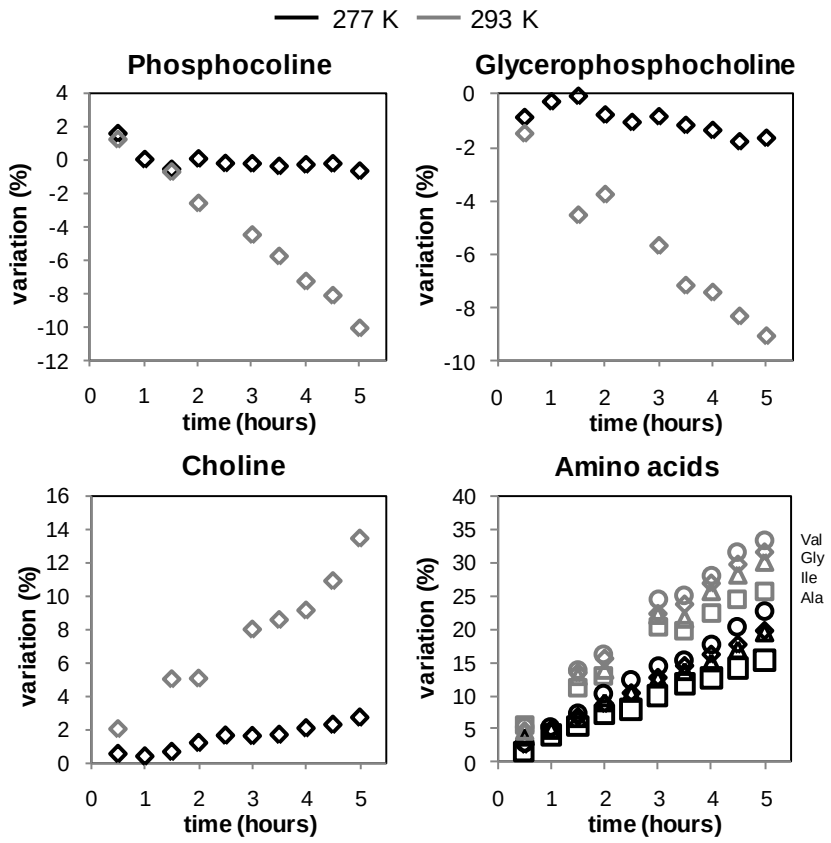


Figure 3.4 Variation (%) of the integrals of choline-containing compounds and amino acids during acquisition of standard 1D ^{1}H HRMAS spectra for 5 hours at SR 4 kHz and two different temperatures, 277 K (black) and 293 K (grey), with respect to the first spectrum (time 0).

Stability at 277 K for up to eight hours (SR 4 kHz) was further verified for three additional tissue samples and the integrals of selected metabolites were measured in spectra recorded at different time points (1.5, 5 and 8 h). The results obtained are expressed in Figure 3.5 as a heatmap colour-coded as a function of the percentage of variation relative to the initial spectrum (time 0 h). Generally, after 1.5 h, metabolite variations were below 10% for all samples, with exception of alanine, glycine and valine. At longer times of sample spinning (up to 8 h), amino acids registered the largest variations, while metabolites like lactate, taurine, *myo*-inositol and PC remained reasonably stable. Overall, these results highlight the importance of controlling sample acquisition time and shows that within the time window used to record the 1D experiments for quantification/ multivariate purposes (1.5 h), tissue stability was fairly maintained, therefore ensuring that spectral profiles were not significantly affected by sample degradation. It should also be noted that 2D spectra with longer acquisition times were exclusively used for assignment purposes.

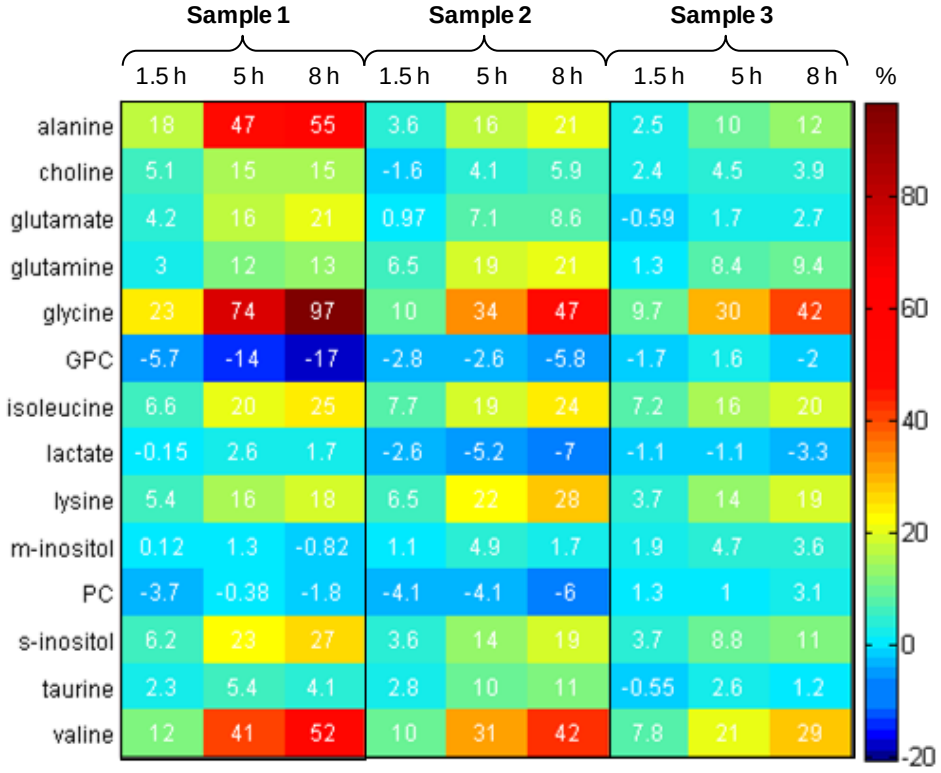


Figure 3.5 Heatmap showing metabolite variations in three tissue samples as function of time in the spectrometer (HRMAS 4 kHz, 277 K). The colour code reflects the percentage of variation relative to the initial spectrum (time 0 h).

3.1.3 Use of relaxation- and diffusion-edited experiments to deal with spectral overlap

As already mentioned, the standard 1D spectra of lung tissues shows the contributions of both small metabolites and macromolecular and/or mobility-restricted components, the latter originating broad resonances and a broad background on the baseline. In order to reduce spectral complexity and overlap, relaxation- and diffusion-edited experiments were recorded to selectively detect small metabolites and macromolecular components, respectively. In particular, the CPMG experiment (explained in section 1.3.2.5) was used to attenuate the signals of fast-relaxing molecules such as lipids (which, as shown in Annex VI, exhibit shorter T_2 time constants). In order to choose the most appropriate echo time ($2\tau n$), a series of tests varying the spin echo delay (τ) and the number of loops (n) was performed. As shown in Figure 3.6, lipid resonances could not be fully attenuated without a concomitant large decrease in small metabolite signals, thus indicating that the lipid moieties detected had relatively high mobility. Using a total echo

time of 90 ms, about 50 % reduction of lipid fatty acyl chain resonances was obtained compared to the standard 1D spectra, enabling the visibility of narrow signals from low M_w metabolites to be improved (Figure 3.7a, b). Therefore, this value was selected for recording all CPMG spectra. Similar values have been reported in the literature for lung and other tissue types (Imperiale et al. 2011; Yang et al. 2013; Benahmed et al. 2013).

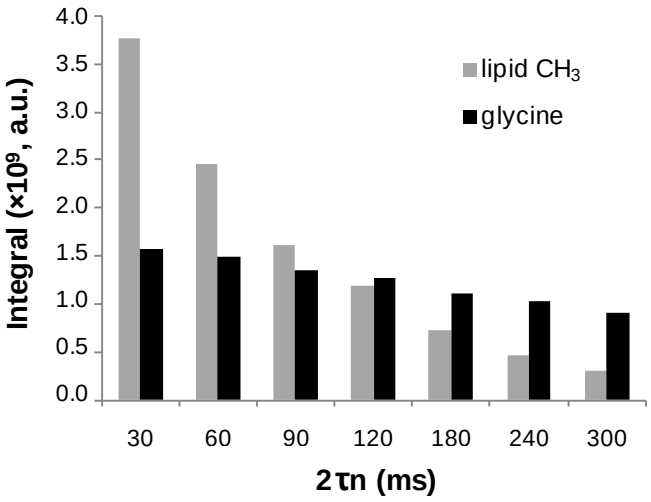


Figure 3.6 Integral areas of lipid (C*H*$_3$) and glycine signals, obtained for different total echo times (2τn) of the CPMG experiment.

Oppositely, the diffusion-edited experiment (described in subchapter 1.3.2.5) selects the signals of slow-diffusing macromolecules, such as lipids and peptides. Under the conditions selected for this experiment – diffusion time of 100 ms and diffusion gradient length of 2 ms, which agree with the values recommended in the literature (Beckonert et al. 2010) – almost all small metabolite signals were attenuated (Figure 3.7c). However, residual signals of lactate and some amino acids (e.g., alanine, taurine) could still be observed, indicating their restricted mobility.

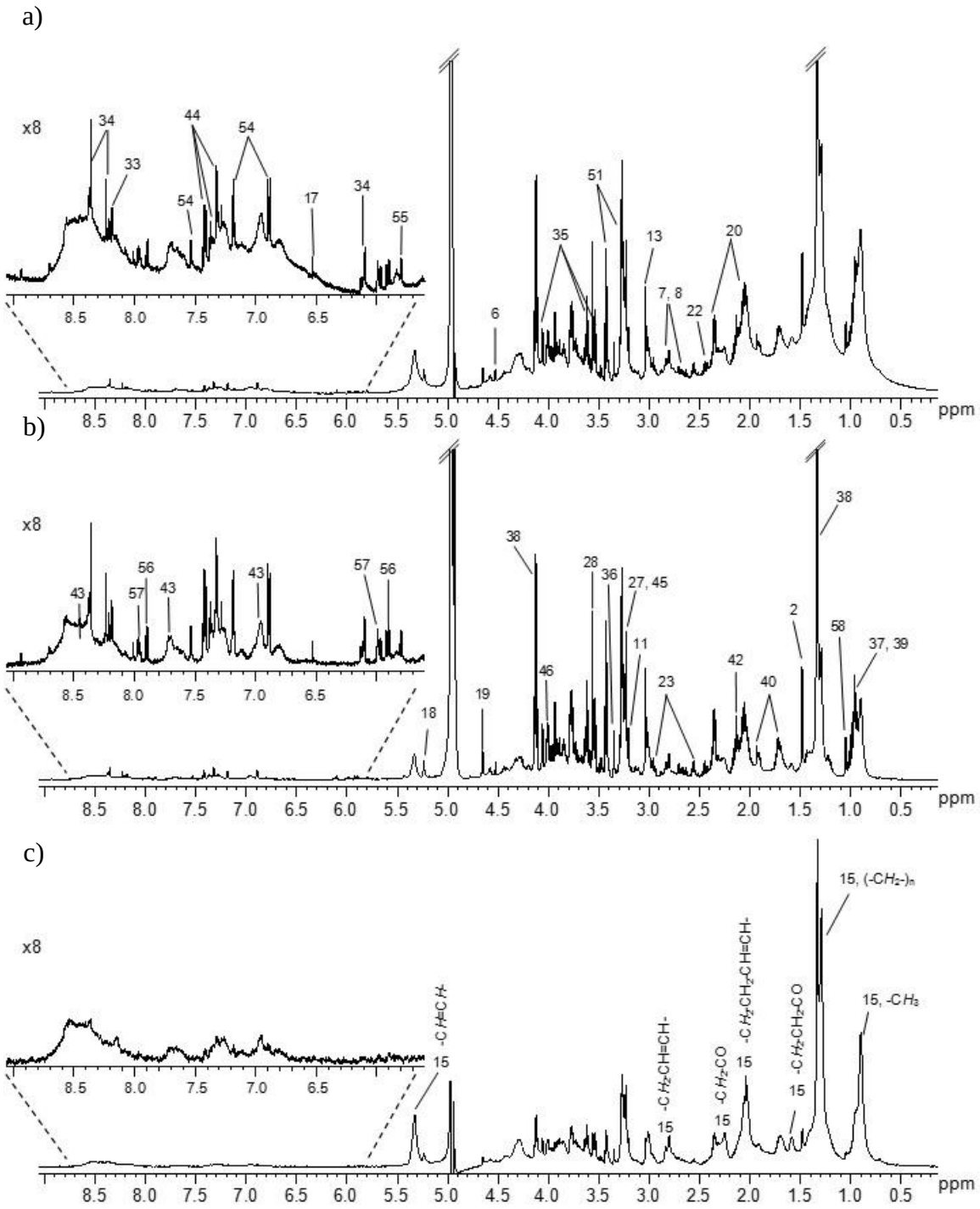


Figure 3.7 Typical 600 MHz ^{1}H HRMAS NMR spectra of lung tumour tissue, used for spectral assignment: a) standard 1D, b) CPMG, c) diffusion-edited. Metabolites are numbered in accordance with Table 3.2.

3.2 Metabolic composition of human lung tissues: spectral assignment based on 1D and 2D NMR experiments

The metabolic composition of intact lung tissue was firstly reported in the literature by our group (Rocha et al. 2010), based on the thorough interpretation of 1D and 2D HRMAS spectra. Besides the complementary three 1D experiments mentioned above, the following 2D experiments have been recorded for selected samples: i) ^{1}H-^{1}H TOCSY, useful for revealing intramolecular spin-spin connectivities; ii) ^{1}H-^{13}C HSQC, helpful to identify singlets or aid the assignment of crosspeaks still overlapped in TOCSY (taking advantage of the larger chemical shift dispersion of ^{13}C resonances); and iii) *J*-resolved, also very helpful in reducing signal overlap and allowing for information on signal multiplicity to be obtained. Figure 3.8 illustrates expansions of typical 2D NMR spectra with some assignments indicated. Table 3.2 shows the complete list of compounds identified, along with their ^{1}H and ^{13}C chemical shifts, measured in 1D and 2D HRMAS spectra.

The broad resonances dominating the low-frequency region (δ 0-3) arise mainly from lipids, as viewed by the characteristic spin systems of fatty acyl chains found in the TOCSY spectrum (Figure 3.8a) and further confirmed by the ^{13}C chemical shifts in the HSQC spectrum (Figure 3.8b). Peptides are also thought to contribute to the broad profile of lung tissues due to the observation of several spin systems (identified in the TOCSY spectrum) with patterns slightly shifted from those of free amino acids; in particular, and in agreement with published data (Tugnoli et al. 2006), the broad signals centred at δ 4.1-4.4, which show spectral correlations to signals in the low-frequency region, are assigned to α-CH protons of different amino acid residues bonded in a peptidic chain (Figure 3.8a and Table 3.2). In addition, some broad resonances in the high-frequency region (δ 6.98, 7.72 and 8.25-8.75) showed strong STOCSY correlations between each other and with this region of α-protons at δ 4.1-4.4, thus suggesting their assignment to backbone and lateral chain NH protons (Figure 3.9a). While these protons are expected to be NMR-invisible in the solution state, due to rapid chemical exchange with water, their detection in tissue indicates a slower exchange rate, possibly in relation with high intracellular viscosity and reduced solvent accessibility.

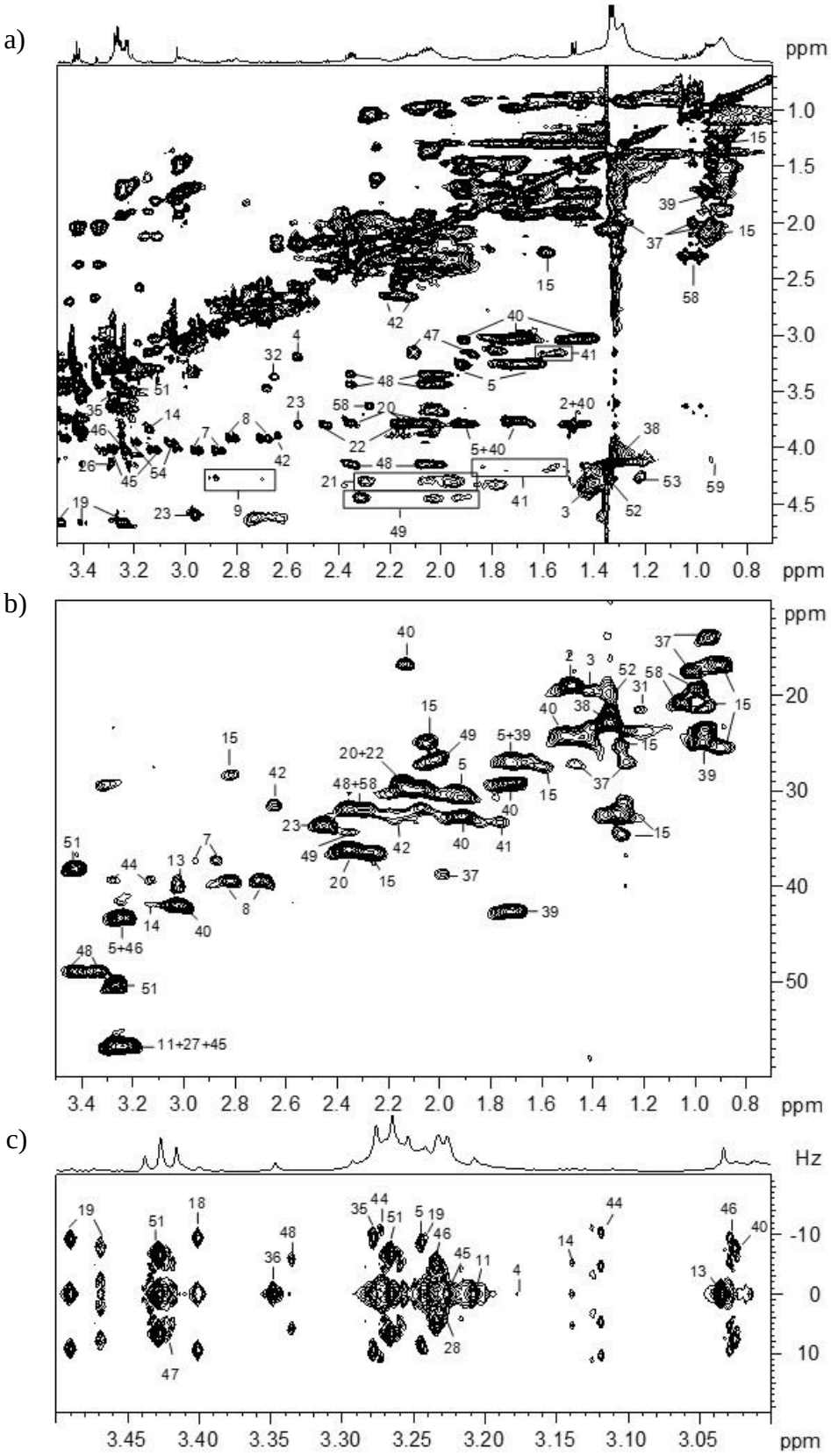


Figure 3.8 Expansions of 2D ^{1}H HRMAS NMR spectra (600 MHz) of lung tumour tissue, used for spectral assignment: a) ^{1}H-^{1}H TOCSY, b) ^{1}H-^{13}C HSQC, c) *J*-resolved. Metabolites are numbered in accordance with Table 3.2.

Table 3.2 Assignment of resonances in the 500/600 MHz ^{1}H NMR spectra of lung tissue; (s, singlet; d, doublet; t, triplet; q, quartet; m, multiplet; dd, double doublet; br, broad). Metabolites which, to our knowledge, are newly reported in this work to be present in human lung tissue are indicated with an asterisk.

No.	Compound	δ ^{1}H in ppm (multiplicity, assignment) / δ ^{13}C in ppm
1	Acetate	1.93 (s, βCH_3)/27.62
2	Alanine	1.48 (d, βCH_3)/19.00; 3.78(q, αCH)/53.41
3	Alanine (bonded)*	1.41 (d, βCH_3)/19.45; 4.31(αCH)
4	β-Alanine*	2.56 (t, βCH_2); 3.18(t, αCH_2)
5	Arginine	1.69 (m, γCH_2)/26.51; 1.92 (m, βCH_2)/30.68; 3.24 (t, δCH_2)/43.31; 3.78 (αCH)/57.39
6	Ascorbate	3.75 (CH_2(OH))/65.38; 4.03 (CH(OH))/72.60; 4.53 (d, C1H)/81.25
7	Asparagine	2.87 (dd, βCH)/37.35; 2.95 (dd, β'CH)/37.35; 4.01 (dd, αCH)/54.25
8	Aspartate	2.69 (dd, βCH)/39.40; 2.82 (dd, β'CH)/39.40; 3.90 (dd, αCH)/55.17
9	Aspartate (bonded)*	2.69 (dd, βCH); 2.89 (dd, β'CH); 4.27 (dd, αCH)
10	Carnitine	2.44 (m, γCH_2); 3.45 (m, αCH_2)
11	Choline	3.21 (s, $N(CH_3)_3$)/56.75; 3.53 (m, CH_2(NH))/70.38; 4.07 (m, CH_2(OH))/58.63
12	Citrate	2.52 (d, α, βCH_2); 2.68 (d, α', $\beta'CH_2$)
13	Creatine	3.03 (s, CH_3)/40.05; 3.94 (s, CH_2)/56.47
14	Ethanolamine	3.14 (t, $CH_2(NH_2)$)/41.91; 3.82 (t, CH_2(OH))/59.35
15	Fatty acyl chain peaks	0.90/0.96 (br,CH_3)/16.85/25.48/21.36; 1.29 (br, $(CH_2)_n$)/25.20/32.42/34.55; 1.58 (br, $\underline{CH_2}$-CH_2-CO)/25.02; 2.04/2.08 (br, CH=CH-CH_2-$\underline{CH_2}$)/27.43; 2.25 (br, CH_2-CO)/36.43; 2.81 (br, CH=CH-$\underline{CH_2}$)/28.31; 5.32 (br, CH=CH)/131.06
16	Formate*	8.46 (s, CH)
17	Fumarate	6.52 (s, CH)
18	α-Glucose	3.40 (t, C4H)/72.48; 3.54 (dd, C2H)/74.10; 3.71 (t, C3H)/75.58; 3.83 (m, C6H)/63.33; 3.85 (m, C5H)/74.85; 5.24 (d, C1H)/94.9
19	β-Glucose	3.24 (dd, C2H/77.09; 3.41 (t, C4H)/72.48; 3.47 (dd, C5H)/78.64; 3.49 (t, C3H)/78.24; 3.74 (dd, C6H)/63.98; 3.90 (m, C6'H)/63.98; 4.65 (d, C1H)/98.78
20	Glutamate	2.06 (m, βCH)/29.87; 2.13 (m, β'CH)/29.87; 2.35 (m, γCH_2)/36.16; 3.78 (t, αCH)/57.38
21	Glutamate (bonded)*	1.95 (βCH); 2.04 (β'CH); 2.30 (γCH_2); 4.30 (αCH)
22	Glutamine	2.15 (m, βCH_2)/ 29.52; 2.45 (m, γCH_2)/ 33.65; 3.79 (t, αCH)/57.38
23	Glutathione (GSH)	2.18 (βCH_2 Glu); 2.56 (γCH_2 Glu); 2.96 (βCH_2 Cys); 3.80 (αCH Gly, αCH Glu); 4.58 (αCH Cys); 8.56 (NH Cys); 8.37 (NH Gly)
24	Glycerol	3.56/3.65 (dd, $C1H_2$/$C3H_2$)/65.29; 3.78 (dd, C2H)/75.12
25	Glyceryl of lipids	4.09, 4.30 (br, $C1H_2$/$C3H_2$); 5.23 (br, C2H)
26	Glycerophosphoetha -nolamine (GPE)*	3.30 (CH2(N)); 4.11 (CH2(P))

Table 3.2 (continued)

No.	Compound	δ ^{1}H in ppm (multiplicity, assignment) / δ ^{13}C in ppm
27	Glycerophospho -choline (GPC)	3.23 (s, N(H3)3)/56.75; 3.71 (β'CH2(N)/68.80; 4.33 (α'CH2(P))/62.29
28	Glycine	3.56 (s, αCH_2)/68.6923
29	Glycogen	3.62 (C2H); 5.43 (C1H)
30	Histidine	7.28 (s, C4H, ring); 8.08 (s, C2H, ring)
31	β-Hydroxybutyrate	1.20 (d, γCH_3)/21.49
32	Hypotaurine	2.65 (t, βCH_2); 3.35 (t, αCH_2)
33	Hypoxanthine	8.18 (s, C2H); 8.21 (s, C8H)
34	Inosine/Adenosine	3.86 (C5',5''H, ribose); 4.28 (C4'H, ribose); 4.44 (C3'H, ribose); 4.78 (t, C2'H, ribose);6.10 (d, C1'H, ribose); 8.23 (s, C2H, ring); 8.36 (s, C8H, ring)
35	*myo*-Inositol	3.28 (t, C5H)/77.16; 3.54 (C1H, C3H)/74.10; 3.62 (C4H, C6H)/75.30; 4.06 (t, C2H)
36	*scyllo*-Inositol	3.35 (s, CH)/76.70
37	Isoleucine	0.94 (t, δCH_3)/13.97; 1.01 (d, $\beta' CH_3$)/17.45; 1.26 (m, γCH)/27.06; 1.47 (m, γ'CH)/27.06; 1.99 (m, βCH)/38.81; 3.65(m, αCH)/62.40
38	Lactate	1.33 (d, βCH_3)/20.89; 4.12 (q, αCH)/71.41
39	Leucine	0.96 (d, δCH_3)/23.64; 0.97 (d, $\delta' CH_3$)/24.93; 1.70 (m, γCH)/27.05; 1.72 (m, βCH_2)/42.65; 3.74 (t, αCH)/56.10
40	Lysine	1.47 (m, γCH_2)/24.28; 1.73 (m, δCH_2)/29.20; 1.92 (m, βCH_2)/32.75; 3.03 (t, εCH_2)/42.00; 3.76 (t, αCH)/57.34
41	Lysine (bonded)*	1.56 (γCH_2); 1.75 (δCH_2)/33.37; 1.83 (βCH_2); 3.16 (εCH_2); 4.17 (αCH)
42	Methionine	2.13 (s, S-CH_3)/16.75; 2.14 (m, βCH)/32.54; 2.21 (m, β'CH)/32.54; 2.64 (t, γCH_2)/31.51; 3.87 (dd, αCH)
43	Peptides*	6.98 (br), 7.72 (br), 8.53 (br)
44	Phenylalanine	3.13 (dd, βCH)/39.40; 3.27 (dd, β'CH)/39.40; 3.99 (dd, αCH)/58.98; 7.32 (m, C2H, C6H, ring)/132.36; 7.37 (m, C4H, ring)/130.42; 7.42 (C3H, C5H, ring)/132.09
45	Phosphocholine	3.22 (s, $N(CH_3)_3$)/56.75; 3.62 (N-CH_2)/69.09; 4.19 (PO_3-CH_2)/60.92
46	Phosphoethanol-amine	3.23 (t, $N(CH_2)$)/57.39; 3.98 (PO_3-CH_2)/63.89
47	Polyamines*	1.79; 2.11; 3.03; 3.14
48	Proline	2.01 (γCH_2)/26.69; 2.07 (βCH)/31.89; 2.35 (β'CH)/31.89; 3.34 (δCH)/48.95; 3.42 (δ'CH)/48.95; 4.13 (αCH)/64.10
49	Proline (bonded)*	1.91 (γCH_2); 2.03 (βCH)/34.39; 2.32 (β'CH)/34.39; 3.69 (δCH)/ 65.70; 3.83 (δ'CH)/65.70; 4.44 (αCH)
50	Serine	3.85 (αCH_2)/59.34; 3.95 (dd, βCH)/63.25; 3.99 (dd, β'CH)/63.25
51	Taurine	3.26 (t, S-CH_2)/50.34; 3.43 (t, N-CH_2)/38.20
52	Threonine	1.34 (d, γCH_3)/19.81; 3.59 (d, αCH)/63.25; 4.26 (m, βCH)/68.72
53	Threonine (bonded)*	1.22 (d, γCH3); 4.26 (d, αCH)

Table 3.2 (continued)

No.	Compound	δ ^{1}H in ppm (multiplicity, assignment) / δ ^{13}C in ppm
54	Tyrosine	3.20 (dd, βCH); 3.05 (dd, β'CH); 3.94 (αCH)/59.26; 6.89 (d, C3H, C5H, ring)/118.66; 7.18 (d, C2H, C6H, ring)/133.75
55	Uracil	5.80 (d, C5H, ring); 7.53 (d, 6H, ring)
56	Uridine	4.13 (C4'H, ribose); 4.23 (C3'H, ribose); 4.35 (C2'H, ribose); 5.89/5.92 (d, C1'H, ribose/C5H, ring) 7.89 (d, C6H, ring)
57	UDP/UTP	5.97 (C1'H, ribose); 7.97 (C6H, ring)
58	Valine	0.99 (d, γCH_3)/19.45; 1.04 (d, $\gamma'CH_3$)/20.75; 2.28(m, βCH)/31.99; 3.62 (d, αCH)/63.34
59	Valine (bonded)*	0.94 (γCH_3); 4.11 (αCH)

Reduced glutathione (GHS), a tripeptide of glutamate, cysteine and glycine, was also unambiguously identified through the TOCSY correlations between α and β protons of glutamate and cysteine, and between α and amide protons of cysteine and glycine. Moreover the STOCSY plot obtained when using the glutamate γCH_2 signal as driver peak (δ 2.56) further confirmed this assignment, as strong correlation was found between all signals attributed to GSH, including the NH signals in the high-frequency region (Figure 3.9b). STOCSY was also useful to corroborate the assignment of glycogen, a polysaccharide of glucose, as, in addition to the TOCSY crosspeak between δ 5.43 (attibuted to C1H of α1→4 linked glucose units) and δ 3.6 (C2H, C4H), strong STOCSY correlations were found between the driver peak at δ 5.43 and signals around δ 3.8 (C5H, C6H) (Figure 3.9c), thus agreeing with literature (Bollard et al. 2000).

In terms of low molecular weight metabolites, more than twenty amino acids were identified, together with some organic acids (e.g., acetate, lactate, citrate, formate, β-hydroxybutyrate), nucleosides/nucleotides, glucose, and inositols (Table 3.2). Several ethanolamine and choline compounds were also detected, the *J*-resolved experiment being particularly useful in this respect, since the overlapping signals in the δ 3.2-3.3 region were clearly separated in the *J*-resolved spectrum (Figure 3.8c). In total, over fifty compounds were identified, together with eight unknown spin systems present in TOCSY and HSQC, thus providing valuable information on the metabolic composition of lung tissue and setting the basis for interpreting the cancer-related variations discussed in the next subchapters.

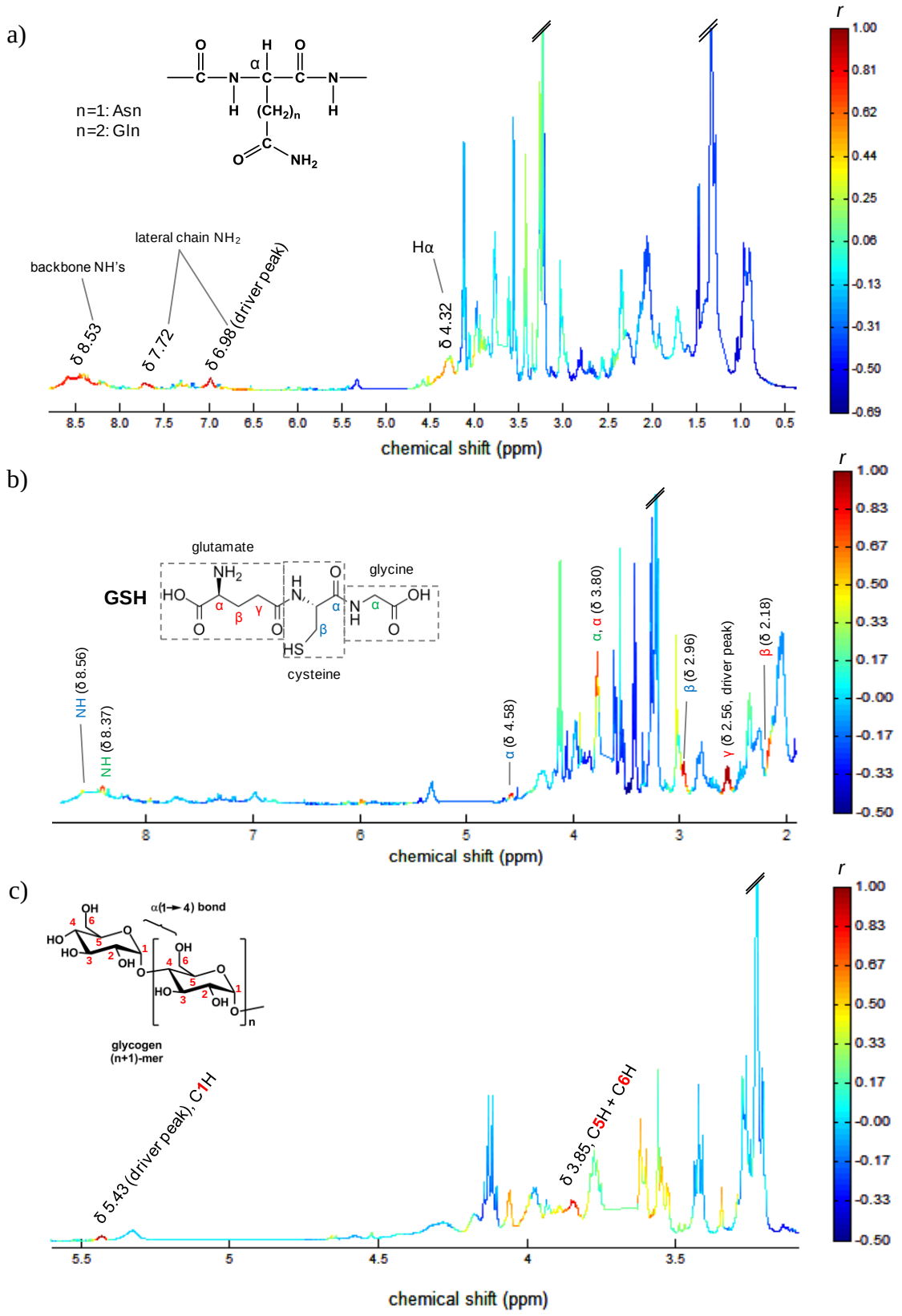


Figure 3.9 STOCSY correlation plots of standard 1D spectra (tumour n 57) of driver peaks: a) δ 6.98, b) δ 2.56 and c) δ 5.43, used for the assignment of small peptides, reduced glutathione (GSH) and glycogen, respectively.

3.3 General metabolic features of lung tumour tissues

The following sections present the NMR metabolic profiles of lung tumour tissues, when compared to control tissues and when taking into account different factors, namely the percentage of tumour cells, necrosis and tumour stage.

3.3.1 Differentiation between tumour and control tissues

Figure 3.10 shows the average ^{1}H HRMAS NMR spectra of lung tumours and their corresponding adjacent non-involved tissues (controls). Visual comparison of the two spectra allowed several differences to be anticipated, namely in the levels of lactate, lipids, phosphocholine (PC), glycerophosphocholine (GPC), creatine, taurine and glutathione (GSH) (increased in tumours), and of acetate and glucose (decreased in tumours). Additionally, the majority of tumours presented higher intensity of broad signals in the aromatic region, probably arising from peptides (as previously discussed).

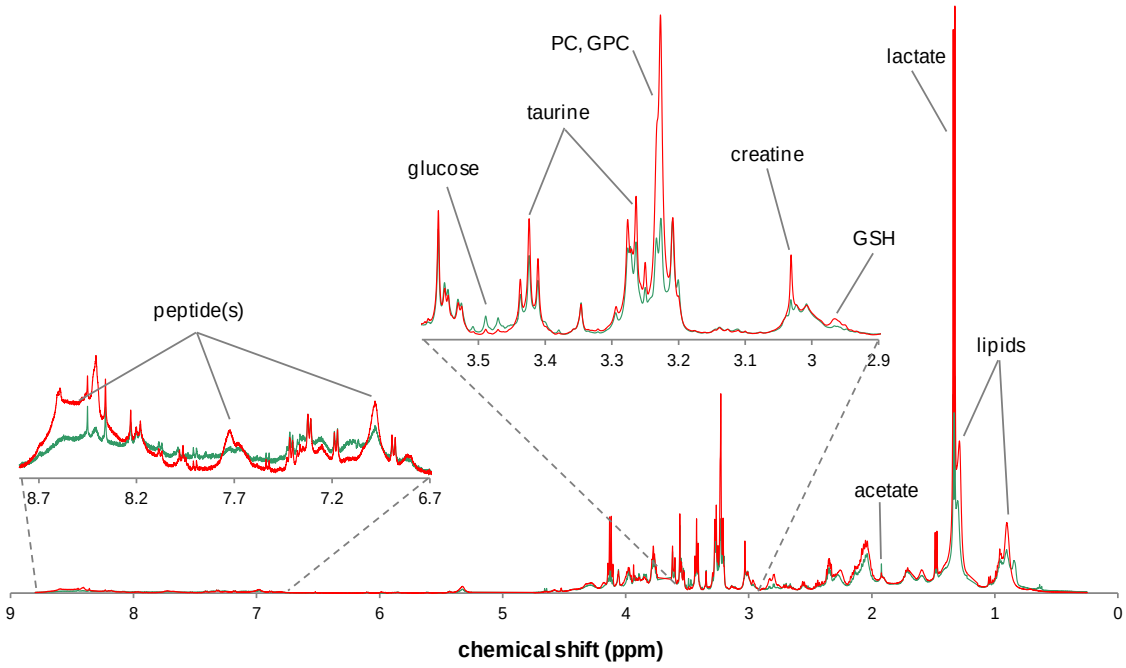


Figure 3.10 Average standard 1D ^{1}H HRMAS NMR spectra of non-involved lung tissues (n 56, green) and lung tumour tissues (n 57, red), after PQN normalisation.

The metabolic features differentiating tumour and control tissues were further investigated by multivariate analysis. Firstly, PCA and PLS-DA were applied to the standard 1D spectra of lung controls (n 57) and tumours (n 58). The resulting PCA scores scatter plot (Figure 3.11a) already shows a trend for separation between the two tissue types, which becomes clearer when PLS-DA is applied (Figure 3.11b).

PLS-DA model robustness was then verified by MCCV and permutation testing, with prediction results being plotted in the ROC space (Figure 3.11c). Whilst the original models (true classes assigned) showed 96.1% sensitivity, 97.8% specificity and an overall classification rate of 96.9% (Table 3.3), random permutation of class membership resulted in a decrease of these parameters to values of no discrimination (approximately 50%). Moreover, the Q^2 values obtained for the original model showed a high median (Q^2 0.83) and a distribution clearly distinct from that obtained for models in which classes were permuted (Figure 3.11d), therefore confirming the predictive power of the original model.

The thorough inspection of PLS-DA LV1 loadings coloured according to variable importance in the projection (VIP) (Figure 3.11e) allowed the main compounds contributing for tissue discrimination to be identified. Lactate, alanine, glutamate, GSH, creatine, taurine, PC, GPC, phosphoethanolamine (PE), uridine di/triphosphate (UDP/UTP) and peptide moieties (with negative loadings) were increased in tumours, whereas acetate and glucose (positive loadings) were decreased.

PCA and PLS-DA were also performed using the CPMG and diffusion-edited experiments. Again, separation between tumour and control tissue was observed and MCCV validated the PLS-DA models obtained, although with inferior classification parameters, especially when using the diffusion-edited data (Table 3.3). The PLS-DA loadings of the CPMG data (not shown) revealed the same variations as observed when using the standard 1D experiment as input for multivariate analysis, although showing lower importance for the partially attenuated broad resonances (namely those in the aromatic region). Conversely, the PLS-DA loadings of the diffusion-edited model (not shown) revealed mainly variations in small metabolites that had not been completely attenuated, such as lactate, taurine, alanine, PC and GPC, not adding further information. Therefore, the standard 1D experiment was selected for further assessment of lung tumours' metabolic profile.

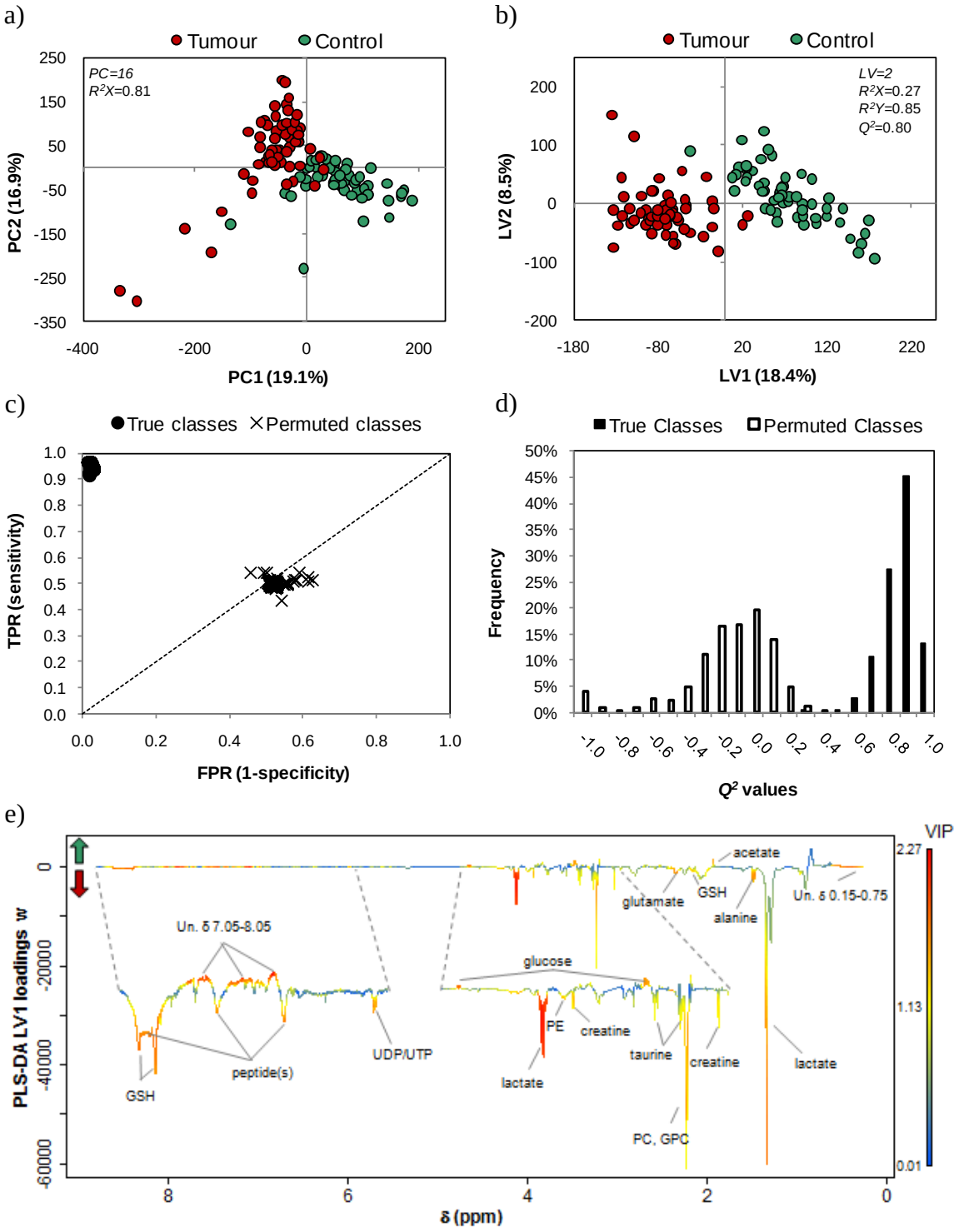


Figure 3.11 MVA applied to the standard 1D ^{1}H HRMAS NMR spectra of lung control (n 57) and tumour (n 58) tissues: a) PCA and b) PLS-DA scores scatter plot. The parameters shown on the scores plot (*PC*: principal component, *LV*: latent variable, R^2X: variation explained by the X matrix, R^2Y: variation explained by the Y matrix, Q^2: predictive power) derive from default 7-fold cross validation. c) ROC space (TPR: true positive rate; FPR: false positive rate) and d) Q^2 histogram obtained by MCCV and permutation testing (500 iterations) of the PLS-DA model. e) PLS-DA LV1 loadings weights coloured as a function of VIP. Metabolites showing VIP>1 are assigned in the plot. (GPC: glycerophosphocholine; GSH: reduced glutathione; PC: phosphocholine; PE: phosphoethanolamine; UDP/UTP: uridine di/triphosphate; Un.: unknown).

Table 3.3 Prediction results obtained by MCCV (500 iterations) of PLS-DA models assessing the discrimination between lung tumour and control tissues.

PLS-DA models	Median Q^2	Sensitivity (%)	Specificity (%)	Classification rate (%)
Standard 1D	0.83	96.1	97.8	96.9
CPMG	0.77	93.9	96.6	95.2
Diffusion-edited	0.55	77.0	87.6	82.3
13 integrals	0.72	95.6	98.6	97.1

The integrals of all metabolite signals with VIP>1 assigned in Figure 3.11e showed statistically significant differences ($p<0.004$, Bonferroni-corrected) and absolute effect sizes greater than 0.5, thus confirming their importance in the discrimination between tumour and control tissues. The percentages of variation for each discriminant metabolite are presented in Table 3.4 and corresponding boxplot representations are shown in Figure 3.12.

Table 3.4 Metabolites showing statistically significant differences between lung tumour and control tissues. For each metabolite, the average percentage and coefficient of variation were obtained by spectral integration of selected signals. Effect sizes and *p*-values are shown, indicating, respectively, the magnitude and statistical significance of the differences. (s: singlet; d: doublet; t: triplet; q: quartet; m: multiplet; br: broad).

Metabolite (δ, multiplicity)	All tumour types (56 pairs)		
	% variation	effect size	*p*-value[a]
Acetate (1.92, s)	-57.0±17.2	-0.88±0.39	3.3×10^{-6}
Alanine (1.48, d)	37.1±5.3	1.1±0.40	1.2×10^{-6}
Creatine (3.93, s)	63.9±7.3	1.2±0.40	1.2×10^{-8}
Glucose (4.65, d)	-46.3±6.8	-1.7±0.43	7.9×10^{-10}
Glutamate (2.35, m)	38.7±3.8	1.6±0.43	5.8×10^{-10}
GPC (3.23, s)	118.4±10.7	1.3±0.41	5.0×10^{-8}
GSH (2.16, m)	42.8±5.4	1.2±0.40	2.3×10^{-9}
Lactate (4.12, q)	95.6±5.1	2.4±0.49	8.2×10^{-11}
PC (3.22, s)	177.4±18.7	0.95±0.39	6.5×10^{-9}
PE (3.98, m)	30.9±4.5	1.1±0.40	2.0×10^{-7}
Taurine (3.42, t)	56.0±7.3	1.1±0.40	6.6×10^{-7}
UDP/UTP (5.96, m)	255.2±17.4	1.2±0.40	6.3×10^{-9}
Peptide (7.72, br)	17.8±5.8	0.54±0.38	2.2×10^{-4}

[a] Wilcoxon signed rank test $p<0.004$ (Bonferroni-corrected).

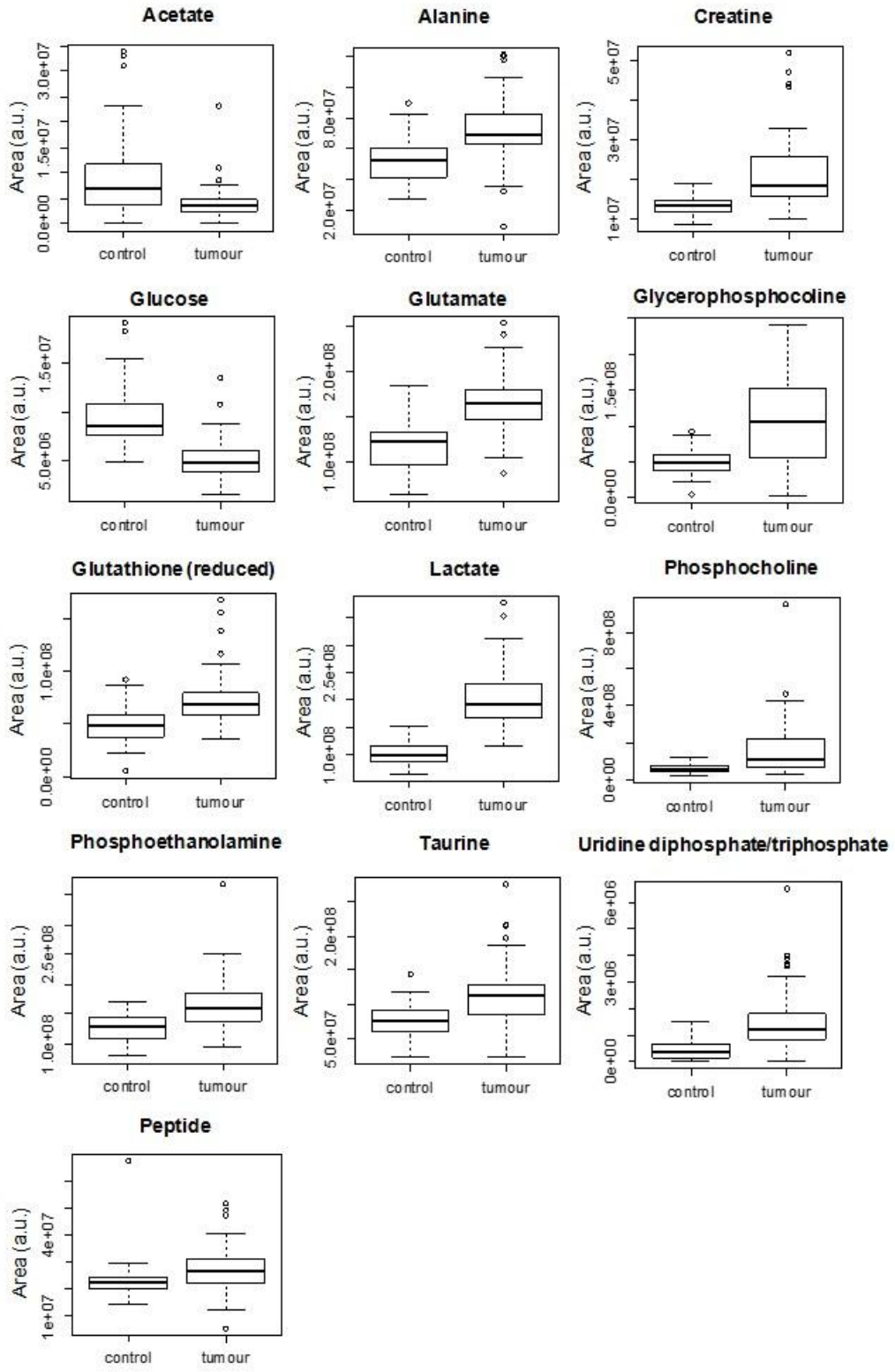


Figure 3.12 Boxplots representing the median, 1st and 3rd quartiles (Q1 and Q3), minimum and maximum of integral values of selected signals (PQN-normalized) for control (n 57) and tumour (n 58) tissues. The points above the maximum [>Q3 + 1.5(Q3 - Q1)] or below the minimum [<Q1 - 1.5(Q3 - Q1)] are outliers.

Further confirmation of the discriminatory power of this panel of metabolites was achieved by performing PLS-DA of their integrals, as model robustness and accuracy were maintained (Table 3.3 and Figure 3.13), with an overall classification rate of 97%.

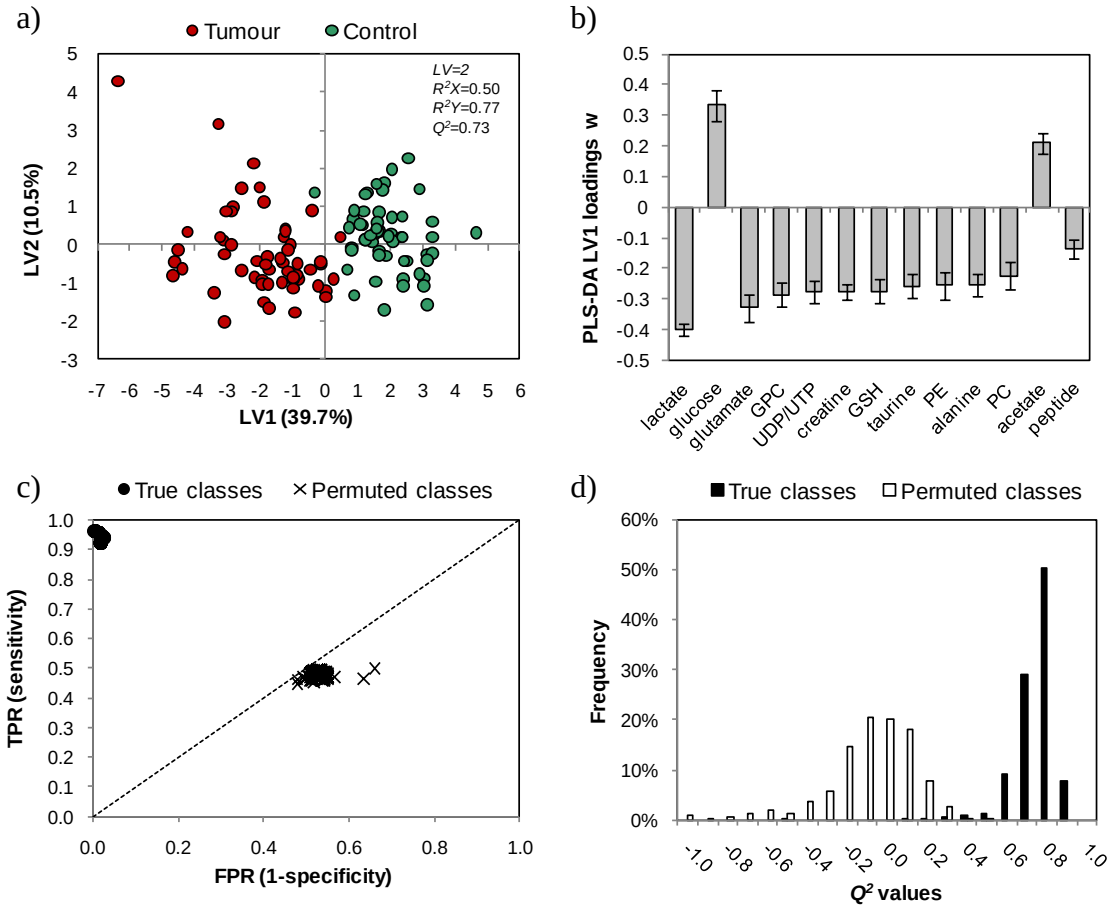


Figure 3.13 PLS-DA applied to 13 selected integrals measured in the standard 1D ^{1}H HRMAS NMR spectra from lung control (n 57) and tumour (n 58) tissues: a) scores scatter plot and b) LV1 loadings. The parameters shown on the scores plot (*LV*: latent variable, R^2X: variation explained by the X matrix, R^2Y: variation explained by the Y matrix, Q^2: predictive power) derive from default 7-fold cross validation. c) ROC space (TPR: true positive rate; FPR: false positive rate) and d) Q^2 histogram obtained by MCCV and permutation testing (500 iterations) of the PLS-DA model. (GPC: glycerophosphocholine; GSH: reduced glutathione; PC: phosphocholine; PE: phosphoethanolamine; UDP/UTP: uridine di/triphosphate).

3.3.2 Impact of the percentage of tumour cells and necrosis on tumour metabolic profile

As the amount of tumour cells on mirror sections of samples analysed by NMR was found to vary between 5 and 99% (Table A6, Annex IV), median and mode values being 50% and 90%, respectively, the possible influence of such variability on tumours'

metabolic composition was addressed by PLS1 regression, considering either the full spectral profile or the set of 13 discriminant metabolites (listed in Table 3.4). The resulting Q^2 values were very low (about 0.1 in both cases), thus indicating that the metabolic profiles of tumour tissues were not strongly determined by the varying proportions of stromal and tumour cells on each sample. Moreover, the possible correlation between the NMR spectral profiles and the amount of necrotic tissue (Table A6, Annex IV) was also investigated through PLS1 regression analysis. The resulting model (n 53) showed a reasonable predictive power (Q^2 of 0.48 for 2 latent variables), with the majority of less necrotic samples located in negative LV1 and samples with higher percentage of necrosis distributed towards positive LV1 (Figure 3.14a). The corresponding VIP-coloured loadings (Figure 3.14b) indicated that higher lipid levels characterised highly necrotic samples, which was confirmed by the significant correlations ($p<0.003$) found between lipid signal areas and % necrosis. Detailed analysis of such correlations further showed the ratio CH=CH to -CH_3 lipid signals (representing the proportion of unsaturated fatty acids) to be only moderately correlated with necrosis (r 0.5, $p<0.05$). Oppositely to what was observed for brain tissues (Sjøbakk et al. 2013), this suggests that no particular change in lipids unsaturation is associated with necrosis.

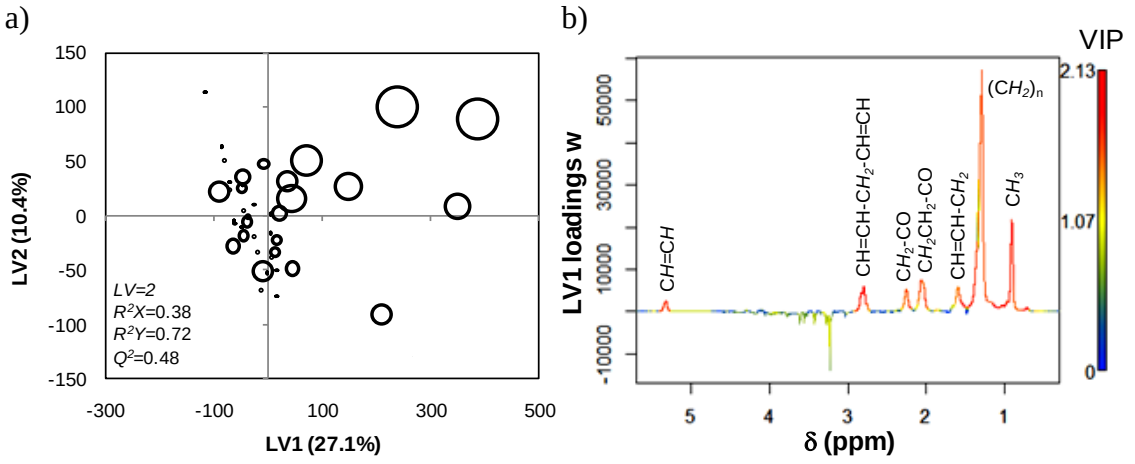


Figure 3.14 PLS1 regression analysis of standard 1D ^{1}H NMR spectra of lung tumour tissues and the amount of necrosis (%) in their respective mirror sections: a) scores scatter plot of the first two latent variables, where the diameter of each circle represents the % of necrosis, b) LV1 loadings plot coloured as a function of VIP (lipid signals are assigned to specific fatty acyl chain protons, in italic). The parameters shown on the scores plot (*LV*: latent variable, R^2X: variation explained by the X matrix, R^2Y: variation explained by the Y matrix, Q^2: predictive power) derive from default 7-fold cross validation.

3.3.3 Impact of stage on tumour metabolic profile

The impact of stage on the tumour profile has been addressed by building separate PLS-DA models for each stage (I: 29 pairs, II: 19 pairs, III: 6 pairs). While the stage III model couldn't be properly validated, likely due to the low number of samples available, the classification rates obtained for discriminating stage I or stage II tumours from their respective controls were similar to the one achieved when considering the whole dataset (Table 3.5).

Table 3.5 Prediction results obtained by MCCV (500 iterations) of tissue PLS-DA models assessing the discrimination between different groups of samples, including different histological types (T: tumour; C: control; AdC: adenocarcinoma; SqCC: squamous cell carcinoma).

PLS-DA classes	Median Q^2	Sensitivity (%)	Specificity (%)	Classification rate (%)
All samples: C (n 57) vs. T (n 58)	0.83	96.1	97.8	96.9
Stage I: C (n 29) vs. T (n 29)	0.83	97.9	95.6	96.7
Stage II: C (n 19) vs. T (n 19)	0.79	94.4	93.3	93.9
AdC: C (n 19) vs. T (n 19)	0.81	86.5	89.1	87.8
SqCC: C (n 19) vs. T (n 19)	0.80	98.6	96.8	97.7
AdC (n 18) vs. SqCC (n 18) (T38 and T56 excluded)	0.47	82.4	80.3	81.3
AdC (n 18) vs. SqCC (n 18) (after variable selection)	0.58	99.2	89.7	94.3
AdC (n 18) vs. SqCC (n 18) (6 signal areas as variables)	0.54	90.6	75.3	82.8

Moreover, inspection of the loadings profiles (not shown) revealed that the main discriminant features were the same for all stages, with the exception of taurine which did not appear important in the discrimination of stage III tumours from their controls. Concordantly, no valid discrimination could be achieved between tumour tissues of different stages, thus indicating that, in the set of samples studied, the metabolic signature of tumour did not change significantly with stage.

3.4 Metabolic features of different tumour histological types

3.4.1 Dependence of tumours' metabolic behaviour on histological type

Figure 3.15 shows the average spectra of control tissue and of each tumour histological type analysed in this study. Some of the differences between tumour profiles are highlighted, suggesting that there may be indeed a relationship between tumour histomorphology and metabolic composition. For instance, carcinoids were characterized by very low lipid levels and high taurine and peptide abundance, whereas small cell carcinomas showed prominently high levels of creatine and *myo*-inositol together with absence of glucose. Regarding the other subtypes, there were also several apparent differences, for instance in the relative levels of creatine, PC, GPC, glucose and glycogen. However, given the low sample numbers representing most types (≤6 patients), only the profiles of adenocarcinomas (AdCs) and squamous cell carcinomas (SqCCs) were further investigated in this study, with two main purposes: i) to find out if these two types have distinct metabolic behaviour, and ii) to evaluate the potential of using NMR metabolomics as an adjunct tool in AdC vs. SqCC differential classification. To address the first goal, the discrimination between each of the two histological types and their respective control tissues was investigated. MCCV-validated PLS-DA models were obtained for both AdC and SqCC compared to controls, with overall classification rates being 87.8 and 97.7%, respectively (Table 3.5). The corresponding scores scatter plots, ROC maps, Q^2 distributions and loadings are shown in Figure 3.16 (a-d for AdC and e-h for SqCC).

Interestingly, VIP-couloured PLS-DA loadings showed that the importance of several metabolites in tumour vs. control discrimination differed between the two types (AdC and SqCC), which was indeed confirmed by spectral integration, as shown in Table 3.6. For instance, adenocarcinomas showed higher increases in PC, GPC, PE, UDP/UTP and peptides, together with a higher decrease in acetate, whereas squamous cell carcinomas were characterized by higher increases of lactate, glutamate, alanine, GSH and creatine, together with a more pronounced decrease of glucose (Table 3.6).

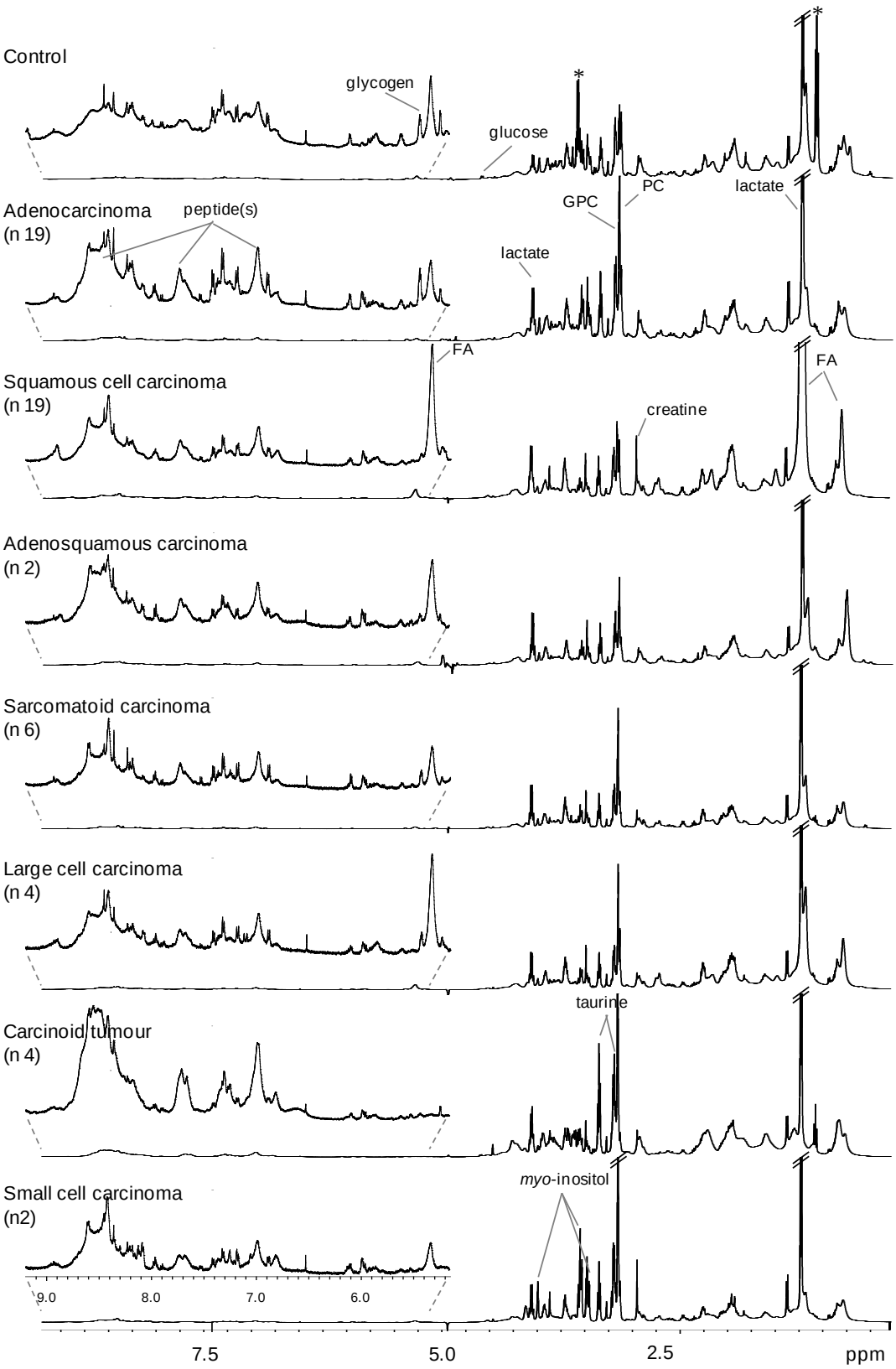


Figure 3.15 Average standard 1D ^{1}H HRMAS NMR spectra of lung control and tumour tissues of different histological types. Some apparent metabolite differences are indicated (FA: fatty acyl chains in lipids, PC: phosphocholine, * ethanol contamination).

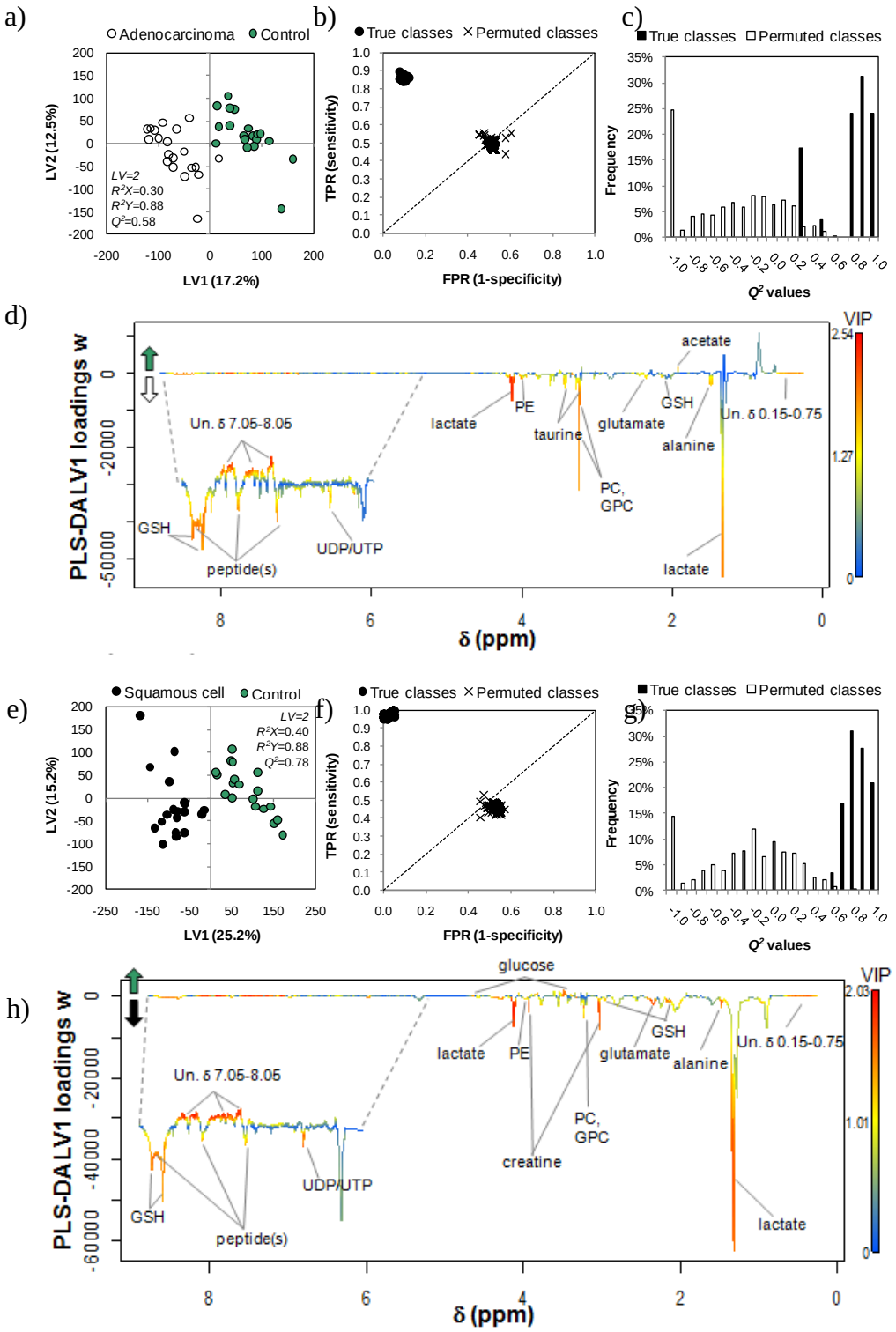


Figure 3.16 Scores scatter plots, ROC spaces, Q^2 histograms and LV1 loadings plots (coloured as a function of VIP) resulting from applying PLS-DA and MCCV (500 iterations) to the standard 1D ^{1}H HRMAS NMR spectra of tumour tissues from squamous cell carcinoma (a-d) or adenocarcinoma (e-h) and their paired controls. The parameters shown on the scores plots (*LV*: latent variable, R^2X: variation explained by the X matrix, R^2Y: variation explained by the Y matrix, Q^2: predictive power) derive from default 7-fold cross validation. (GPC: glycerophosphocholine; GSH: reduced glutathione; PC: phosphocholine; PE: phosphoethanolamine; UDP/UTP: uridine di/triphosphate; Un.: unknown).

Table 3.6 Metabolites showing statistically significant differences between lung tumours and control tissues, considering all samples (2nd column), only adenocarcinomas (3rd column) and only squamous cell carcinomas (4th column). For each metabolite, the average percentage and coefficient of variation were obtained by spectral integration of selected signals. Effect sizes and *p*-values are shown, indicating, respectively, the magnitude and statistical significance of the differences. (s: singlet; d: doublet; t: triplet; q: quartet; m: multiplet; br: broad; GPC: glycerophosphocholine; GSH: reduced glutathione; PC: phosphocholine; PE: phosphoethanolamine; UDP/UTP: uridine di/triphosphate).

Metabolite (δ, multiplicity)	**All tumour types (56 pairs)**			**Adenocarcinoma (19 pairs)**			**Squamous cell (19 pairs)**		
	% variation	**effect size**	***p*-value[a]**	**% variation**	**effect size**	***p*-value[a]**	**% variation**	**effect size**	***p*-value[a]**
Acetate (1.92, s)	-57.0±17.2	-0.88±0.39	3.3×10^{-6}	-61.9±22.2	-1.3±0.70	1.3×10^{-4}		n.s.	
Alanine (1.48, d)	37.1±5.3	1.1±0.40	1.2×10^{-6}		n.s.		46.9±7.7	1.6±0.73	1.6×10^{-4}
Creatine (3.93, s)	63.9±7.3	1.2±0.40	1.2×10^{-8}	28.2±6.7	1.2±0.69	6.4×10^{-4}	126.4±12.1	2.1±0.79	3.8×10^{-6}
Glucose (4.65, d)	-46.3±6.8	-1.7±0.43	7.9×10^{-10}	-29.3±10.8	-1.0±0.68	2.0×10^{-3}	-51.4±11.4	-2.0±0.77	3.8×10^{-6}
Glutamate (2.35, m)	38.7±3.8	1.6±0.43	5.8×10^{-10}	24.2±5.8	1.2±0.69	4.2×10^{-4}	55.8±6.1	2.3±0.82	3.8×10^{-6}
GPC (3.23, s)	118.4±10.7	1.3±0.41	5.0×10^{-8}	115.2±15.8	1.5±0.72	2.1×10^{-4}		n.s.	
GSH (2.16, m)	42.8±5.4	1.2±0.40	2.3×10^{-9}		n.s.		72.5±9.9	1.7±0.75	3.8×10^{-6}
Lactate (4.12, q)	95.6±5.1	2.4±0.49	8.2×10^{-11}	84.8±7.4	2.6±0.86	3.8×10^{-6}	121.8±9.1	2.7±0.88	3.8×10^{-6}
PC (3.22, s)	177.4±18.7	0.95±0.39	6.5×10^{-9}	216.3±21.5	1.6±0.73	1.6×10^{-4}	37.5±11.6	0.88±0.67	2.0×10^{-3}
PE (3.98, m)	30.9±4.5	1.1±0.40	2.0×10^{-7}	32.5±5.8	1.6±0.73	5.2×10^{-4}		n.s.	
Taurine (3.42, t)	56.0±7.3	1.1±0.40	6.6×10^{-7}		n.s.			n.s.	
UDP/UTP (5.96, m)	255.2±17.4	1.2±0.40	6.3×10^{-9}	422.2±40.3	1.1±0.68	8.0×10^{-4}	243.7±25.3	1.4±0.71	1.3×10^{-4}
Peptide (7.72, br)	17.8±5.8	0.54±0.38	2.2×10^{-4}	28.0±9.2	0.87±0.67	2.0×10^{-3}		n.s.	

[a] Wilcoxon signed rank test $p<0.004$ (Bonferroni-corrected). n.s. not significant.

Examination of dependencies between metabolite variations further complemented this analysis, the results being plotted as correlation heatmaps colour-coded by the strength of Spearman correlation coefficients (Figure 3.17). A striking observation was that the inter-metabolite correlation pattern was clearly different between adenocarcinomas and squamous cell carcinomas. In AdC, significant correlations ($r_S>|0.7|$, $p<0.001$) involved mainly phospholipid-related metabolites (PC, GPC and PE), while the SqCC pattern was characterized by correlations involving lactate, glucose, glutamate, alanine, GSH and creatine.

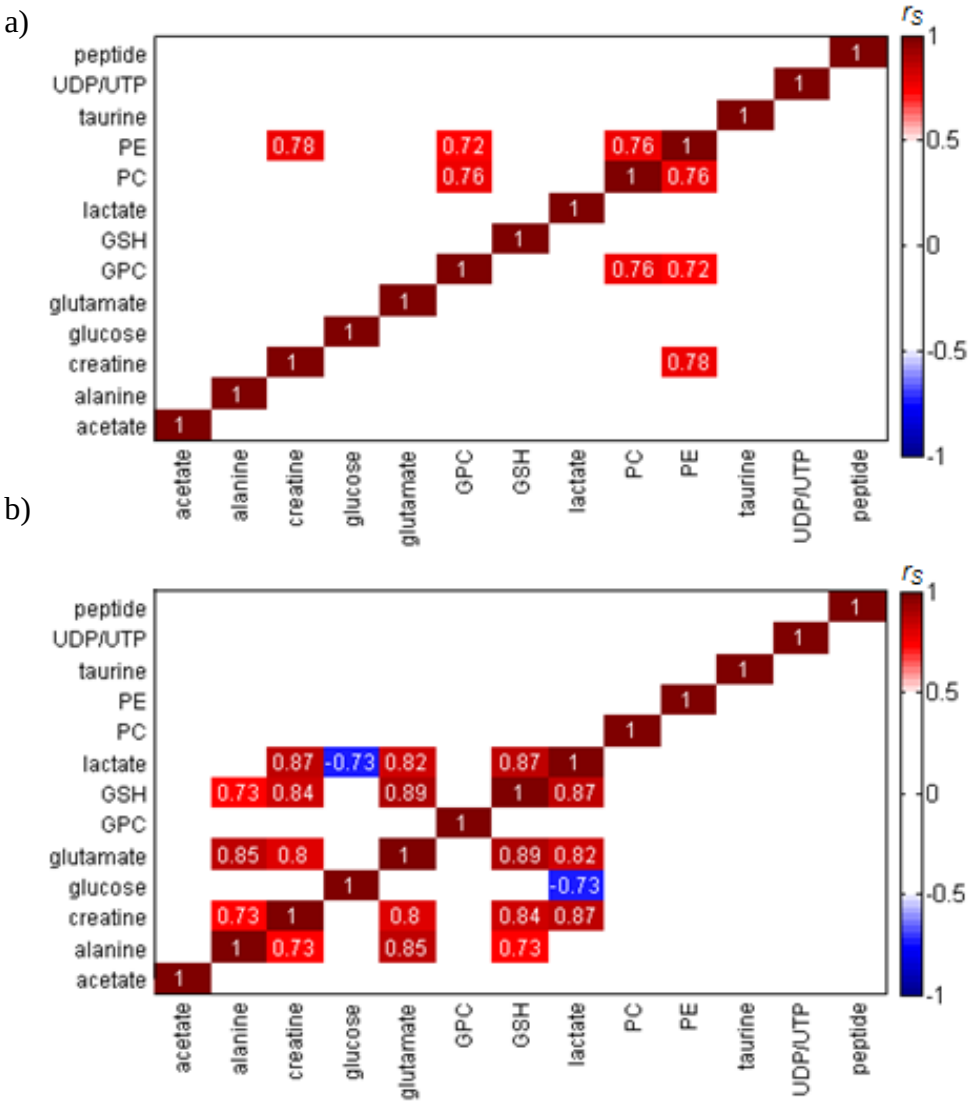


Figure 3.17 Spearman correlation (r_S) heatmaps between metabolites found to be important in tumour vs. control discrimination: a) adenocarcinoma and b) squamous cell carcinoma. A cut-off value of r_S>|0.7| and significance of p<0.001 was applied for visualization purposes.

A selected set of graphs showing the quantitative relationship between metabolites found to be highly correlated in either AdC or SqCC are presented in Figure 3.18. Regarding the levels of GPC or PE vs. PC, significantly correlated in AdCs, the graphs in Figure 3.18a and 3.18b reveal that, in spite of the observed scattering, these relationships did show a difference between most AdC tumours compared to their respective controls, whereas in SqCC the overlap between groups was much higher. In regard to significant correlations in SqCC, the negative correlation between glucose and lactate levels was

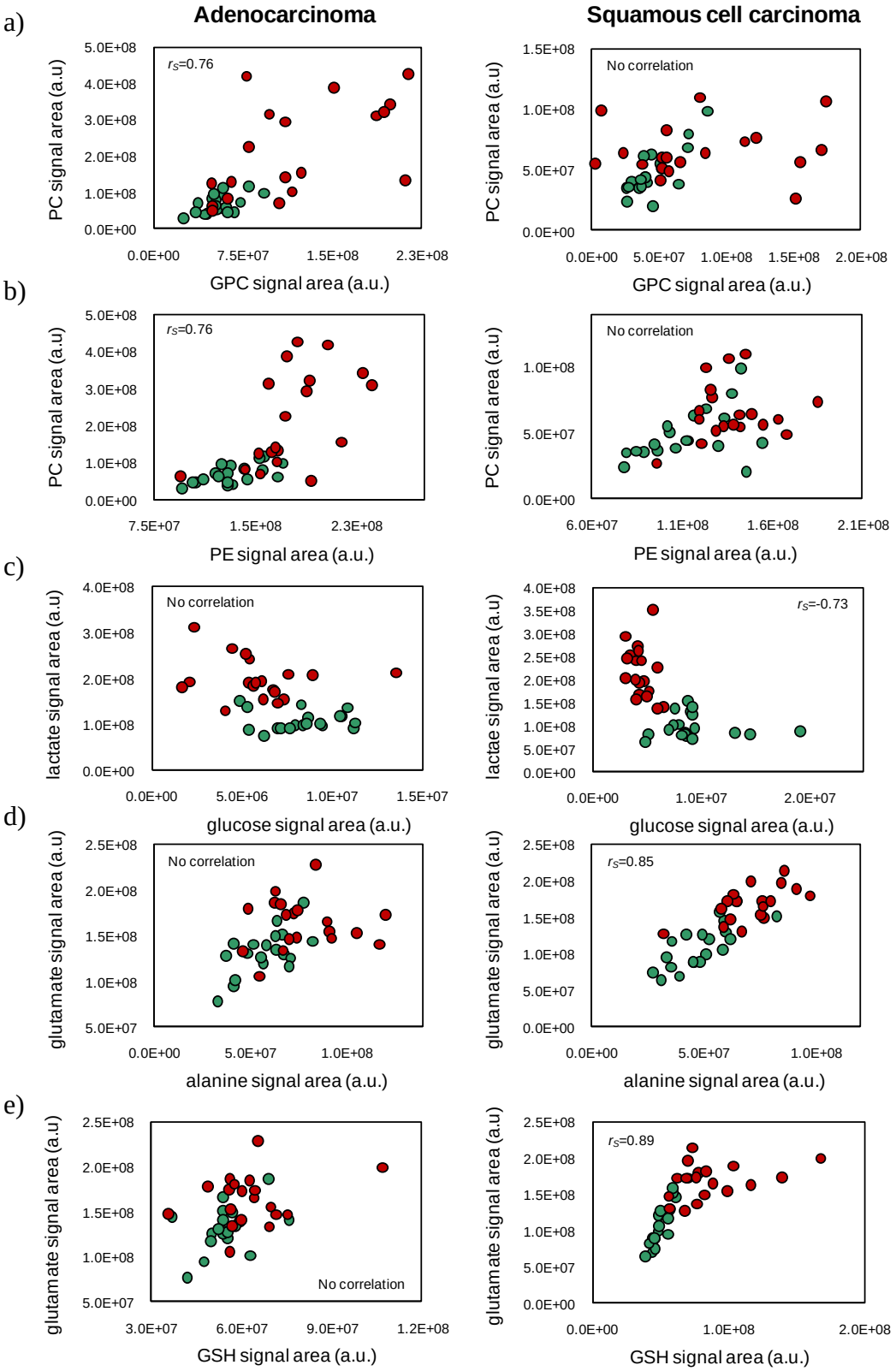


Figure 3.18 Graphical representation of integral areas for pairs of metabolites showing significant Spearman correlations (p<0.001) in either adenocarcinoma (left) or squamous cell carcinoma (right), considering control (green) and tumour (red) tissues: a) GPC vs. PC, b) PE vs. PC, c) glucose vs. lactate, d) alanine vs. glutamate, e) GSH vs. glutamate.

confirmed for tumour samples and found to be less evident in control tissues (Figure 3.18c). Also, in SqCC (but not in AdC), alanine and glutamate were positively correlated and their relationship followed a very clear linear trend which captured group separation (Figure 3.18d), a similar behaviour being observed for the glutamate vs. GSH relationship (Figure 3.18e).

3.4.2 Potential for differentiating adenocarcinoma from squamous cell carcinoma

For assessing the potential of metabolomics to discriminate between AdC and SqCC, multivariate analysis has been applied to their tumour tissue spectra, leaving out control tissues. After excluding two clear outliers (AdC T38 and SqCC T56 characterized by very high lipids), the resulting PLS-DA model showed moderate accuracy and predictive power (classification rate 81.3% and median Q^2 0.47), as determined by MCCV (Table 3.5). This discrimination could, however, be significantly improved by applying the variable selection method described in the subchapter 2.3.5, allowing a model with 94.3% classification rate and median Q^2 0.58 to be built. The corresponding PLS-DA scores plot and MCCV validation results are shown in Figure 3.19.

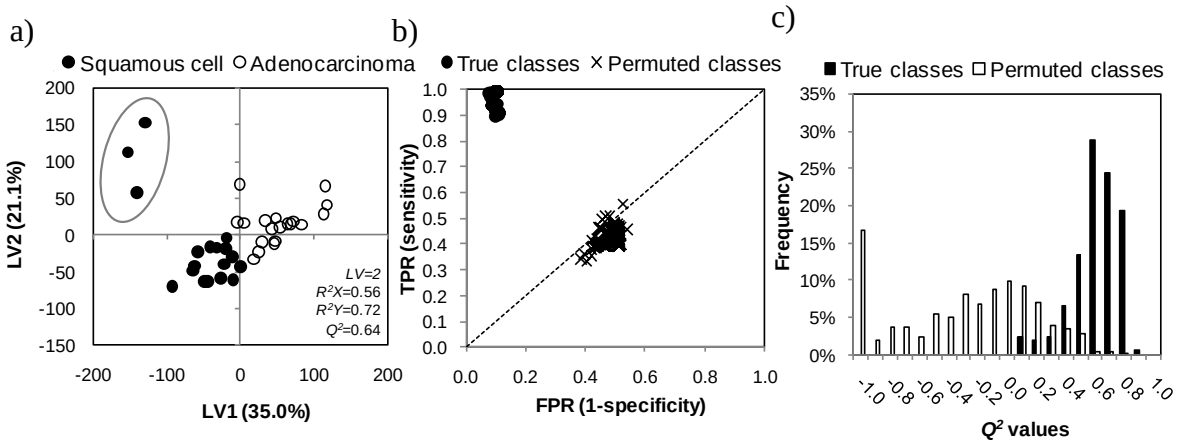


Figure 3.19 a) Scores scatter plot of PLS-DA applied to the standard 1D ^{1}H HRMAS NMR spectra from AdC (n 18) and SqCC (n 18) tissues after variable selection. The parameters shown on the scores plot (*LV*: latent variable, R^2X: variation explained by the X matrix, R^2Y: variation explained by the Y matrix, Q^2: predictive power) derive from default 7-fold cross validation. b) ROC space (TPR: true positive rate; FPR: false positive rate) and c) Q^2 histogram obtained by MCCV and permutation testing (500 iterations) of the PLS-DA model.

In the scores scatter plot, most AdC and SqCC samples are separated along LV1, with three SqCC samples clustering further away, likely in relation with their higher lipid content. Furthermore, because the AdC and SqCC groups compared were unbalanced in terms of gender (AdC: 7M/11F and SqCC: 16M/2F), possible gender-related bias has been assessed. The PCA scores plot, where a trend for separation of the two types was clear, showed no evidence of gender-related clustering (Figure 3.20a, b).

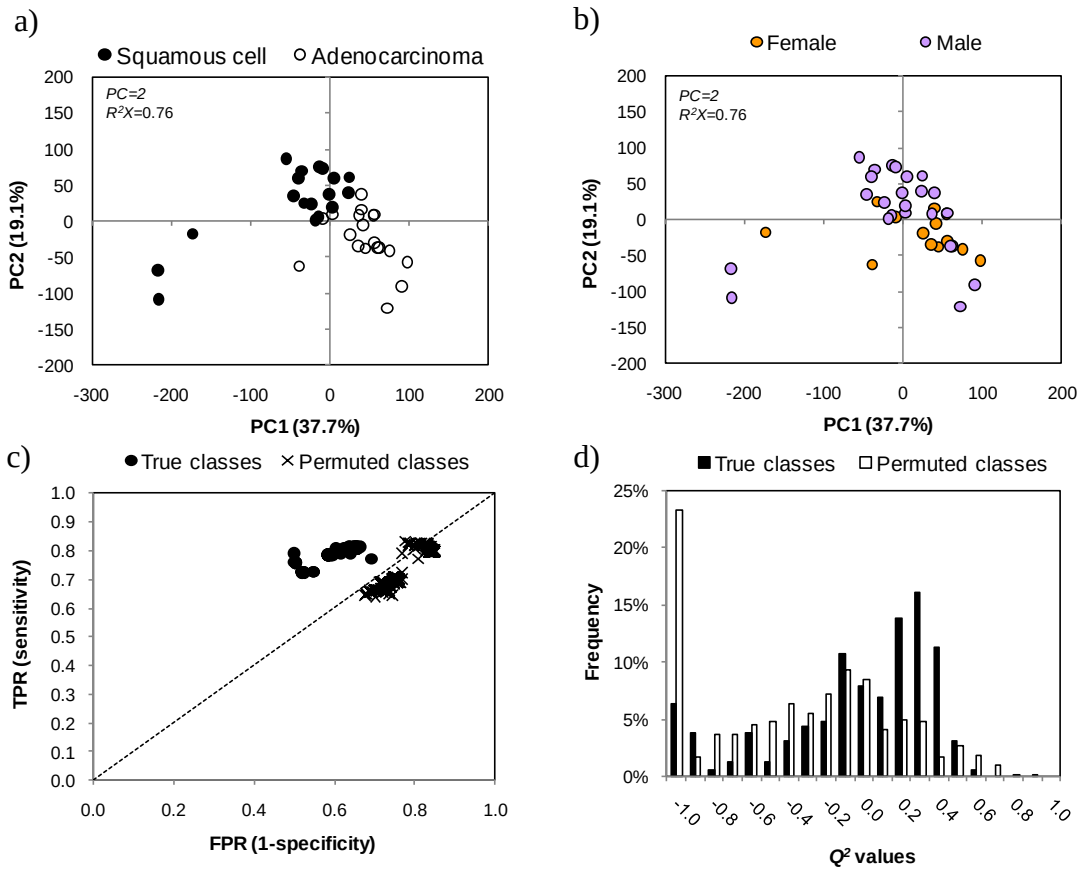


Figure 3.20 PCA scores scatter plot of standard 1D ^{1}H HRMAS NMR spectra from adenocarcinoma (n 18) and squamous cell carcinoma (n 18) tissues, coloured according to a) histological subtype and b) gender. The parameters shown on the scores plots (*PC*: latent variable, R^2X: variation explained by the X matrix) derive from default 7-fold cross validation. c) ROC space (TPR: true positive rate; FPR: false positive rate) and d) Q^2 histogram obtained by MCCV and permutation testing (500 iterations) of the PLS-DA model of male vs. female (adenocarcinoma only).

Moreover, when considering only adenocarcinoma samples (7M/11F), no valid discrimination could be achieved between males and females, as shown by the superimposed Q^2 distributions and ROC spaces of true and permuted models (Figure 3.20c, d). Therefore, gender could be excluded as an important confounding factor. In regard to

age, the difference between groups was not statistically significant (median age 64 for both AdC and SqCC groups), thus reducing the possibility of any age-related bias. Moreover, the subjects' age was found to have a negligible impact on the tumour tissue profile as expressed by the low Q^2 values (<0.3, expressing the cross-validated variability related to age) of PLS regression models built for each of the two major subtypes (not shown).

The metabolic features discriminating adenocarcinoma and squamous cell carcinoma tumour tissues were then inspected by integrating a number of signals within the variables selected. Those showing statistically significant differences ($p<0.002$, Bonferroni-corrected) between AdC and SqCC are listed in Table 3.7, along with the percentages of variation and effect sizes.

Table 3.7 Metabolites showing statistically significant differences between adenocarcinoma and squamous cell carcinoma tumour tissues. For each metabolite, the average percentage and coefficient of variation were obtained by spectral integration of selected signals. Effect sizes and *p*-values are shown, indicating, respectively, the magnitude and statistical significance of the differences. Positive/negative variations refer, respectively, to metabolites increased/decreased in squamous cell carcinomas compared to adenocarcinomas. (s: singlet, t: triplet, m: multiplet, GSH: reduced glutathione, PC: phosphocholine, PE: phosphoethanolamine).

Metabolite (δ, multiplicity)	% variation	effect size	*p*-value[a]
Creatine (3.93, s)	55.8±11.7	1.3±0.71	6.1×10^{-4}
GSH (2.16, m)	36.8±9.8	1.1±0.69	7.0×10^{-4}
myo-Inositol (3.62, t)	-54.5±13.5	-1.8±0.76	8.5×10^{-6}
PC (3.22, s)	-71.6±21.2	-1.7±0.75	2.0×10^{-6}
PE (3.98, m)	-22.4±4.6	-1.8±0.76	3.1×10^{-6}
Taurine (3.43, t)	-29.0±10.1	-1.1±0.69	1.9×10^{-3}

[a] Wilcoxon signed rank test $p<0.002$ (Bonferroni-corrected).

Creatine and GSH were found to be relatively higher in SqCC, whereas PC, PE, *myo*-inositol and taurine were higher in AdC. Interestingly, the later two metabolites, known to play a role in cellular osmoprotection, were not significantly different between each histological type and their respective control tissues. The ability of this set of six metabolites to discriminate between AdC and SqCC was further evaluated by PLS-DA of their signal integrals (Figure 3.21). Separation between the two histological types was maintained in the corresponding scores scatter plot and the loadings confirmed the

variations previously noted (Figure 3.21a, b). A valid model could still be obtained (Figure 3.21c, d), although with lower classification rate (82.8%, Table 3.5), thus showing the importance of considering a more comprehensive profile. Overall, these results pose an opportunity to explore new adjunct methods for the differential diagnosis of the two histological tumour types with highest incidence.

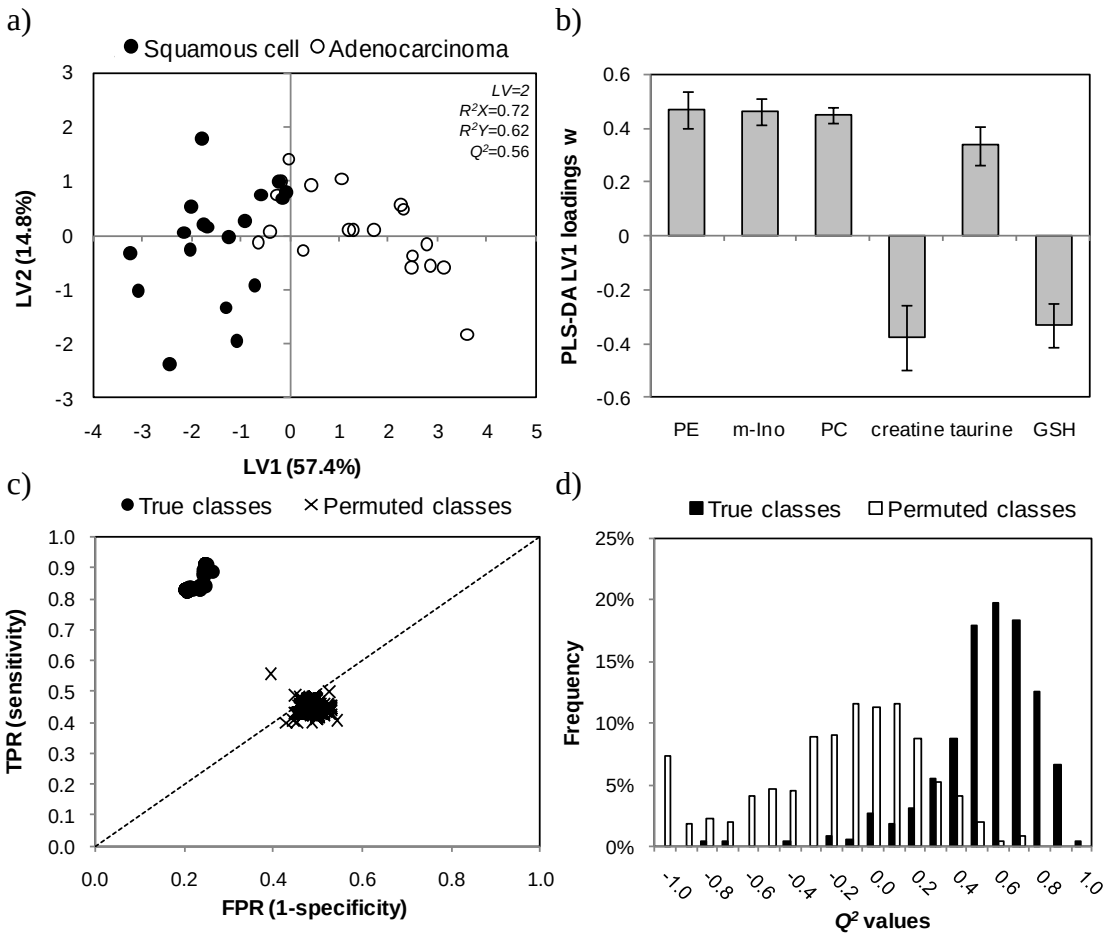


Figure 3.21 PLS-DA of 6 selected signal integrals measured in the standard 1D ^{1}H HRMAS NMR spectra from lung adenocarcinoma (n 18) and squamous cell carcinoma (n 18) tissues: a) scores scatter plot, b) LV1 loadings. The parameters shown on the scores plot (*LV*: latent variable, R^2X: variation explained by the X matrix, R^2Y: variation explained by the Y matrix, Q^2: predictive power) derive from default 7-fold cross validation. c) ROC space (TPR: true positive rate; FPR: false positive rate) and d) Q^2 histogram obtained by MCCV and permutation testing (500 iterations) of the PLS-DA model. (GSH: reduced glutathione; m-Ino: *myo*-inositol; PC: phosphocholine; PE: phosphoethanolamine).

3.5 Proposed biochemical interpretation of tumour-related metabolic changes

The results in the preceding sections provide clear evidence of a strong metabolic signature for lung tumours, mainly defined by a set of thirteen metabolites with significantly altered levels in relation to control pulmonary tissue. Moreover, this signature was found to depend on the tumour histological type, showing important differences between adenocarcinoma and squamous cell carcinoma.

One of the main features characterizing all tumours consisted of increased lactate levels and depleted glucose levels (Figure 3.22), corroborating the expected increase in glycolytic activity ('Warburg effect') reported in the literature for lung tumours (Fan et al. 2009; Kami et al. 2013), as well as for other tumour types, including hepatocellular, prostate and colorectal cancers (Yang et al. 2007; Tessem et al. 2008; Chan et al. 2009). Interestingly, this variation was found in this work to be more pronounced in SqCC than in AdC, glucose and lactate showing a significant negative correlation in the former. These results highlight the dependency of the known glycolytic phenotype of lung tumours, exploited in cancer diagnosis and staging through ^{18}fluorodeoxyglucose (FDG) positron emission tomography (PET) imaging (Ambrosini et al. 2012), on the histological subtype. Concordantly, previous studies have shown that, compared to AdC, SqCC tumors were characterized by higher FDG uptake and increased overexpression of key membrane glucose transporters, namely GLUT1, necessary to support the high rate of glycolysis (Brown et al. 1999; De Geus-Oei et al. 2007; Meijer et al. 2012).

In addition to lactate, glutamate and alanine were also significantly increased in lung tumours, when compared to their paired non-involved tissues. Glutamate may originate from hydrolysis of glutamine (glutaminolysis), which, in addition to glucose, is regarded as an important source of energy and intermediate building blocks for tumour cell growth and proliferation (Wise et al. 2008; Mohamed et al. 2014). Subsequent transamination between glutamate and pyruvate, catalysed by alanine transaminase, also produces alanine and α-ketoglutarate, the latter entering the TCA cycle (Figure 3.22). Thus, the observed variations in glutamate and alanine, also reported in studies of tissue extracts (Fan et al. 2009; Kami et al. 2013), suggest increased glutaminolytic activity in tumour tissues. The strong positive correlation between the two metabolites in SqCC tumours corroborates this hypothesis. In the case of AdC, however, glutaminolysis appears to be less pronounced, as

glutamate registered a smaller increase and alanine did not differ significantly between control and tumour tissues. Glutamate may also leave the glutaminolytic pathway and be excreted, as an effective way to excrete hydrogen, or to act as immunosuppressive in the protection of tumour cells (Mazurek 2007).

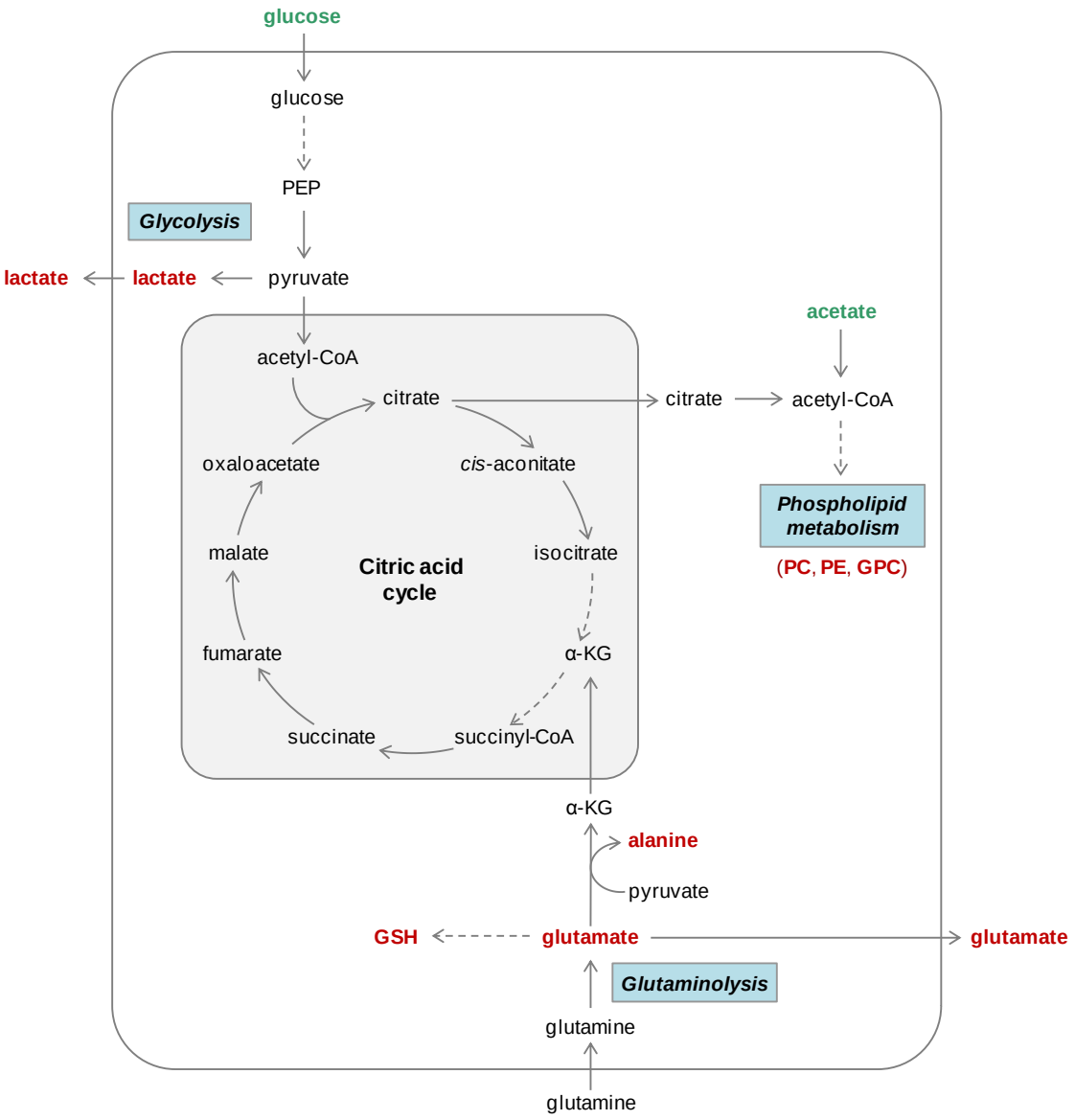


Figure 3.22 Overview of some possibly altered metabolic pathways in lung tumours. Metabolites found to be increased/decreased in tumour tissues relatively to control pulmonary parenchyma are highlighted in red/green.

Moreover, together with cysteine and glycine, glutamate is used for the synthesis of reduced glutathione (GSH, γ-Glu-Cys-Gly), a key component of the cells' defence system for detoxification of oxidative stress-causing species or events. In this work, GSH was

found to be increased in tumour tissues relatively to controls, this variation being more pronounced in SqCC and non-significant in AdC. High intracellular GSH levels together with increased activity of GSH-related enzymes are a typical feature of several tumours (Gamcsik et al. 2012), including lung tumours (Blair et al. 1997; Ferruzzi et al. 2003; Ilonen et al. 2009), and have been associated with chemo-resistance and high metastatic potential (Traverso et al. 2013). Regarding GSH variation in lung tumours of different histology, our findings are consistent with an early work (Cook et al. 1991), but disagree with another study reporting higher GSH increase in AdC than in SqCC, compared to non-malignant adjacent tissue (Ilonen et al. 2009). It should however be noted that other patient characteristics, such as age and smoking habits, are expected to strongly influence the oxidative burden, hence the tissue GSH levels. Therefore, data on the subject's smoking history, which is lacking in our study, would be necessary to further interpret the observed discrepancy and achieve solid conclusions.

Another change that could be related to antioxidant protection mechanisms is the increase in taurine, since, among other physiological roles (e.g., osmoregulation, calcium signalling), this metabolite is known to have important antioxidant activity, recently postulated to be linked to mitochondrial protein synthesis (Jong et al. 2012). According to our results, taurine levels were significantly increased when all tumours were considered together in comparison to controls, although not reaching statistical significance when AdC or SqCC tumours were considered individually. Moreover, it is interesting to note that, together with GSH (but in opposite sense), taurine accounted for the differentiation between AdC or SqCC tumours, thus suggesting that the variations in this metabolite may eventually reflect different antioxidant mechanisms in tumours of different histological types.

Creatine was found to be significantly increased in all tumours, this variation being more prominent in SqCC than in AdC (126 vs. 28% increase, respectively). Elevated creatine levels in tumour tissues is in agreement with previous findings in lung tissue extracts (Yokota et al. 2007) and have been related to altered energetic transfer processes (Somashekar et al. 2011; Yang et al. 2013), which may reflect decreased activity of creatine kinase (CK). Indeed, CK activity has been previously reported to be lower in both AdC and SqCC lung tumours, compared to normal adjacent tissues, although not showing a significant difference between both subtypes (Joseph et al. 1997).

In regard to phospholipid-related metabolites, PC, GPC and PE were found to be increased in tumours relatively to controls. Although, to our knowledge, this has been newly reported for lung cancer by our group, altered phospholipid metabolism has often been linked to cancer as a means to support accelerated cellular proliferation (Sitter et al. 2006; Yang et al. 2007; Tessem et al. 2008; De Silva et al. 2009). PC may arise from both anabolic and catabolic transformations of phophatidylcholine (PtdCho), the major cell membrane phospholipid. A key enzyme involved in this cycle is choline kinase (ChoK), which phosphorylates choline to produce PC, as the first step in PtdCho synthesis. ChoK has been found overexpressed and highly active in lung tumors and cell lines (Molina et al. 2002; Molina et al. 2007), and has been proposed as a prognostic factor for early-stage NSCLC (Molina et al. 2007). Also, ^{11}C-choline PET is being explored as an alternative to ^{18}FDG-PET in lung cancer staging, based on increased choline uptake and metabolism by rapidly proliferating tumor cells (Li et al. 2013). On the other hand, phospholipid breakdown has also been linked to abnormal metabolism in cancer and suggested to promote malignant transformation via mitogenic signal transduction (Moestue et al. 2011). As GPC is solely a product of PtdCho degradation, it is likely that the observed increases of PC and GPC, together with increased PE levels, reflect both anabolic and catabolic processes, which our results suggest to be more relevant in AdC. Additionally, significantly decreased acetate levels in AdC (not observed in SqCC) may also relate to lipid metabolism as this metabolite can be converted to acetyl-CoA to feed *de novo* biosynthesis of lipids. Indeed, the shift from citrate oxidation to lipogenesis is another typical feature of cancer metabolism (Costello and Franklin 2005). However, although total lipids did show a (non-significant) trend to be higher in tumours, their levels seemed to be influenced by other factors like necrosis and did not prove to be determinant in the discrimination between malignant and non-malignant tissue.

Another interesting finding of this work was the increased contribution of peptide moieties to lung tumours' metabolic profiles, in particular the AdC profile. Concordantly, Wu et al. have found elevated levels of 90 dipeptides in lung AdC tissues (from 9 patients) relatively to normal tissues (Wu et al. 2013). In another study, 17 dipeptides were found to be increased in cancer-associated fibroblasts (CAFs) isolated from non-small cell tumours, compared to matched normal fibroblasts obtained from non-neoplastic tissues (Chaudhri et al. 2013). The authors postulated that the observed increase could relate to enhanced

autophagy, a lysosome-related protein degradation mechanism, which has recently emerged as a key regulator of multiple aspects of cancer biology (Kimmelman 2011; Singh and Cuervo 2011).

Finally, uracil nucleotides (UDP and/or UTP) were found to be significantly increased in tumours, especially in AdC. Given the important role of these molecules as extracellular signalling molecules (Morrone et al. 2003) and in the cancer-related glycosylation of proteins and lipids (Christiansen et al. 2014), this is an interesting clue to further investigate their biological relevance in the context of lung cancer.

4 NMR METABOLOMIC STUDY OF BLOOD PLASMA TO ASSESS METABOLIC ALTERATIONS RELATED TO LUNG CANCER

This chapter addresses the ^{1}H NMR analysis of blood plasma from lung cancer patients and healthy controls. After a first section describing the NMR assignment of plasma metabolites, the discrimination between patients and controls is attempted through multivariate analysis. The metabolic differences between the two groups are highlighted, addressing also the dependency of the observed variations on the tumour histological type and stage. Moreover, the possible confounding influence of the subjects' gender, age and smoking habits on the classification models built is assessed, and preliminary external validation is explored. Then, in the last section, a possible biochemical interpretation for the putative cancer-related metabolic variations is presented.

4.1 Metabolic composition of human blood plasma: spectral assignment based on 1D and 2D NMR experiments

Blood plasma is a rich fluid matrix constituted by suspended proteins, particularly albumin, immunoglobulins, glycoproteins and lipoproteins, together with a variety of inorganic and low-molecular weight (Mw) organic solutes. This complex, multi-component nature of plasma results in a ^{1}H NMR profile characterized by a broad envelope of protein signals (mainly lipoproteins) overlapping with narrow signals from small metabolites, as shown by the standard 1D spectrum in Figure 4.1a. In order to maximize the information retrieved from both macromolecular components and small molecules, in addition to the standard 1D spectrum, relaxation- and diffusion-edited experiments have also been recorded for all plasma samples.

As described in subchapter 1.3.2.5, the CPMG experiment enables the attenuation of broad protein resonances based on their reduced molecular mobility, hence, shorter T_2 relaxation time constants (Table A9, ANNEX VI). The resulting CPMG spectrum (Figure 4.1b) showed a substantially flatter baseline and improved visibility of low Mw signals.

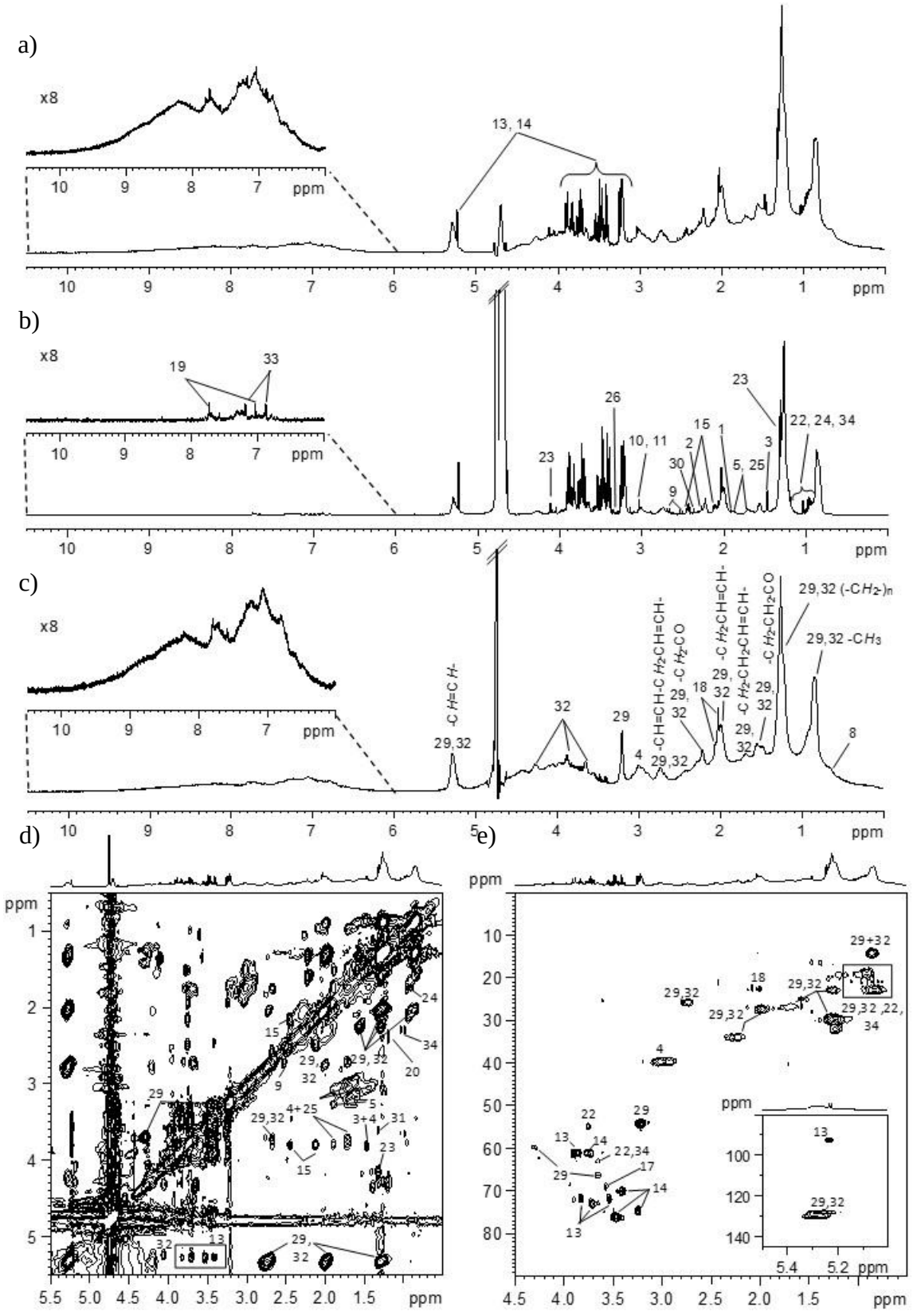


Figure 4.1 Typical 500 MHz ^{1}H NMR spectra of blood plasma of a lung cancer patient, used for spectral assignment: a) standard 1D, b) CPMG, c) diffusion-edited, d) ^{1}H-^{1}H TOCSY and e) ^{1}H-^{13}C HSQC. Metabolites are numbered in accordance with Table 4.1.

For instance, in the aromatic region of the spectrum, the broad protein NH resonances were fully attenuated, allowing for low abundance metabolites like tyrosine and histidine to stand out. Regarding the signals of fatty acyl chains in lipoproteins (e.g., broad resonances at δ 0.8-0.9 and δ 1.2-1.3), their attenuation was not complete (thus indicating their relatively high mobility), but it was enough to greatly improve the visibility of amino acids and other small metabolites in the lower frequency region of the spectrum.

On the other hand, the diffusion-edited experiment filtered off the sharp signals of small metabolites (Figure 4.1c), based on their faster diffusion coefficient (as explained in section 1.3.2.6), allowing for the broad macromolecular profile of blood plasma to be inspected. While the dominant resonances in this profile belong to lipid moieties (mainly composing lipoproteins), albumin lysyl residues, *N*-acetyl groups of glycoproteins and protein NH groups also contributed to the plasma broad envelope.

Two-dimensional NMR experiments, namely ^{1}H-^{1}H TOCSY, ^{1}H-^{13}C HSQC and *J*-resolved, were acquired for selected samples, further helping the characterization of plasma metabolic composition. In particular, the TOCSY spectrum (expansion shown in Figure 4.1d) was very helpful in identifying the spin systems of several amino acids (e.g., leucine, valine, alanine, lysine, glutamine, threonine, tyrosine), small organic acids (e.g., lactate, citrate), as well as in distinguishing α- and β-glucose resonances, and in identifying the homonuclear correlations within the fatty acyl chains of lipoproteins, the glyceryl backbone and the choline-containing phospholipids. Furthermore, the HSQC spectrum (Figure 4.1e) provided ^{13}C chemical shift information which corroborated some of those assignments and enabled others to be proposed. For instance, the ^{1}H-^{13}C correlations for the signals at 2.9-3.0 ppm were in agreement with the assignment to albumin lysyl residues, according to the previously published spin systems of human serum albumin (HSA) (Harris et al. 1996), while those for the singlets at 2.03 and 2.07 ppm matched *N*-acetyl groups of mobile carbohydrate side-chains (mostly *N*-acetylglucosamine and *N*-acetylneuraminic acid) of glycoproteins, as suggested by Bell et al. (Bell et al. 1987).

Statistical correlation spectroscopy (STOCSY) was also applied to the plasma 1D spectra for assignment purposes and proved especially useful for assigning broad resonances to specific lipoprotein subclasses, strongly superimposed even in 2D spectra. As shown in Figure 4.2, the driver peak at δ 0.84 was correlated with broad signals at δ 1.23, 1.49, 1.96, 2.66-2.80, 3.21 and 5.26 (Figure 4.2a), while the resonance at δ 0.87 was

found to be strongly correlated with peaks at δ 1.27, 1.57, 2.00, 2.22, 2.70-2.85 and 5.29, and with glyceryl backbone resonances (δ 4.05, 4.28 and 5.20) (Figure 4.2b). Also, correlations with different cholesterol resonances ($C(18)H_3$) were obtained in each case. Based on the average typical composition of the major lipoprotein subclasses (Nelson and Cox 2004) and on literature data focused on the study of serum/plasma lipoproteins by ^1H NMR (Daykin et al. 2001; Liu et al. 2002; Mallol et al. 2013), the first fraction has been assigned to HDL and the second to VLDL (with contribution from LDL). Indeed, HDL lipoproteins contain large amounts of phospholipids (about 25%), which is consistent with the strong correlation to choline headgroup signals, likely arising from phosphatidylcholine (Figure 4.2a).

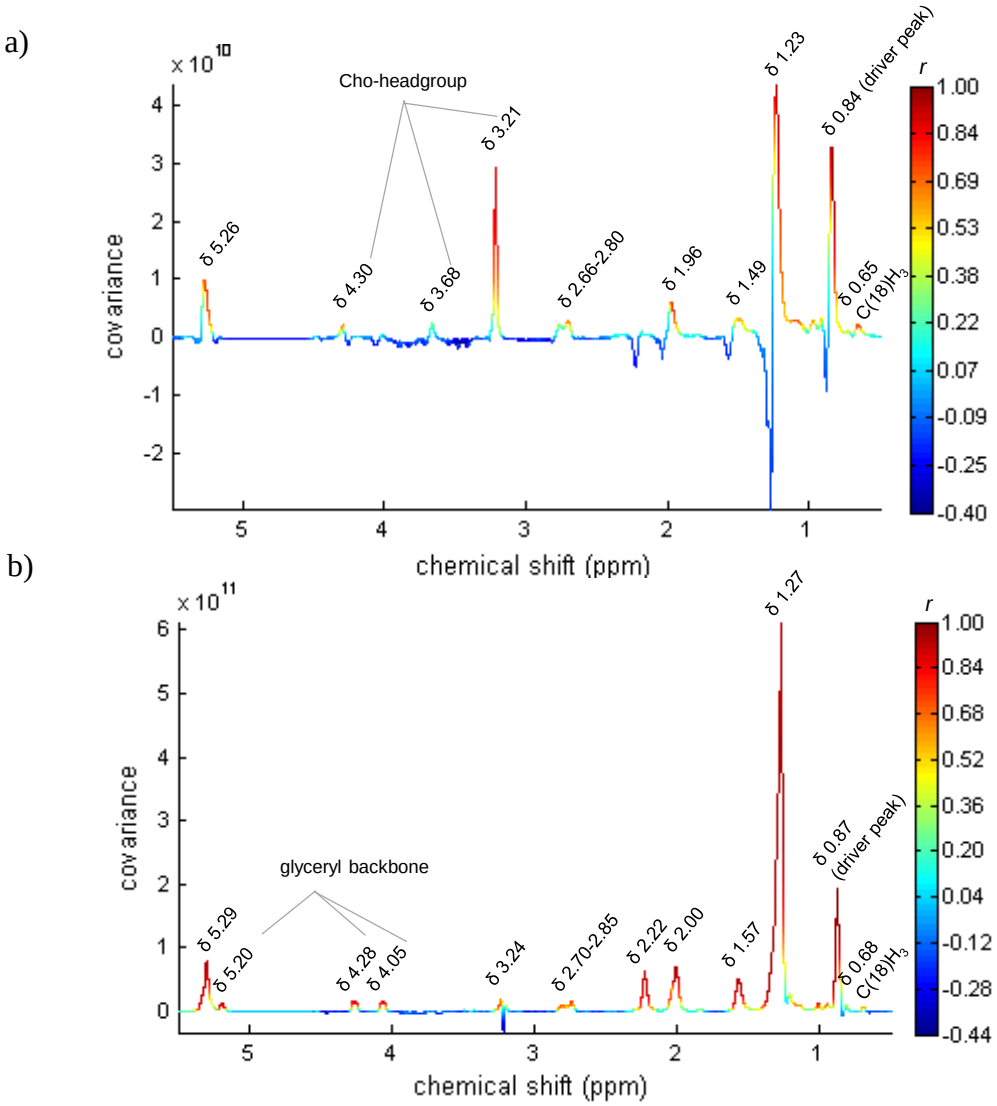


Figure 4.2 STOCSY correlation plots of diffusion-edited spectra (cancer n 106) of driver peaks: a) δ 0.83 and b) δ 0.87.

On the other hand, VLDL lipoproteins include a large proportion of triglycerides in their composition (about 50%), thus agreeing with the spectral profile shown in Figure 4.2b. This difference in the chemical shifts of fatty acyl chains associated with different lipoproteins was also visible in the ^{1}H-^{13}C HSQC, particularly with respect to the unsaturated carbons (expansion in Figure 4.1e).

Spiking with standard solutions was also used in some instances to confirm the assignment of low intensity singlets for which no TOCSY or HSQC crosspeaks were detected (e.g., methanol). The complete assignment of plasma resonances is shown in Table 4.1. Overall, based on the detailed assignment information available in the literature (Nicholson et al. 1995) and on consultation of spectral databases, namely Bruker's BBIOREFCODE-2-0-0 database (Bruker Biospin, Rheinstetten, Germany) and other available on-line databases (Wishart et al. 2007; Psychogios et al. 2011; Ulrich et al. 2008), thirty four compounds have been identified in blood plasma (all previously reported), while seven signals remained unassigned.

Table 4.1 Assignment of resonances in the 500 MHz ^{1}H NMR spectra of blood plasma; (s, singlet; d, doublet; t, triplet; q, quartet; m, multiplet; dd, double doublet; br, broad).

No.	Compound	δ ^{1}H in ppm (multiplicity, assignment) / δ ^{13}C in ppm
1	Acetate	1.91 (s, βCH_3)
2	Acetoacetate	2.22 (s, γCH_3)
3	Alanine	1.47 (d, βCH_3)/16.09; 3.78(q, αCH)
4	Albumin lysyl	2.90 (t, ε-CH_2)/39.42; 2.96 (t, ε-CH_2)/39.42; 3.01 (t, ε-CH_2)/39.59
5	Arginine	1.68 (m, γCH_2)/26.52; 1.88 (m, βCH_2); 3.24 (t, δCH_2)/42.87; 3.76 (αCH)/54.68
6	Asparagine	2.83 (dd, βCH); 2.94 (dd, β'CH)
7	Aspartate	2.67 (dd, βCH); 2.82 (dd, β'CH)
8	Cholesterol	0.65 (br, $C18H_3$ (mainly HDL)); 0.68 (br, $C18H_3$ (mainly LDL+VLDL))
9	Citrate	2.52 (d, α, βCH_2); 2.68(d, α', β'CH_2)
10	Creatine	3.03 (s, CH_3); 3.94 (s, CH_2)
11	Creatinine	3.04 (s, CH_3); 4.05 (s, CH_2)
12	Formate	8.45 (s, CH)
13	α-Glucose	3.40 (t, C4H)/69.69; 3.53 (dd, C2H)/71.65; 3.71(t, C3H)/72.87; 3.82 (m, C6H)/ 71.60; 3.87 (m, C5H)/60.88; 5.23 (d, C1H)/92.16
14	β-Glucose	3.24 (dd, C2H/74.36; 3.40 (t, C4H)/ 69.69; 3.47 (dd, C5H)/76.07; 3.49 (t, C3H)/ 76.07; 3.74 (dd, C6H)/ 60.91; 3.90 (m, C6'H)/ 60.88; 4.65 (d, C1H)/ 96.16
15	Glutamine	2.12 (m, βCH_2)/ 26.74; 2.44 (m, γCH_2)/ 30.99; 3.77 (t, αCH)

Table 4.1 (continued)

No.	Compound	δ ^{1}H in ppm (multiplicity, assignment) / δ ^{13}C in ppm
16	Glycerol	3.62 (dd, $C1H_2/C3H_2$); 3.76 (dd, C2H)
17	Glycine	3.55 (s, αCH_2)/68.69
18	Glycoproteins (*N*-acetyl groups)	2.03, 2.07 (br s, $NHCOCH_3$)/22.33
19	Histidine	7.05 (s, C4H, ring); 7.76 (s, C2H, ring)
20	β-Hydroxybutyrate	1.19 (d, γCH_3)/21.49; 2.35 (m, $\alpha\alpha' CH_2$)
21	Isobutyrate	1.17 (d, CH_3)
22	Isoleucine	0.93 (t, δCH_3)/13.97; 1.00 (d, $\beta' CH_3$); 1.26 (m, γCH)/27.87; 1.40 (m, γ'CH)/27.87; 1.96 (m, βCH); 3.65(m, αCH)/62.83
23	Lactate	1.32 (d, βCH_3)/22.83; 4.11 (q, αCH)/68.83
24	Leucine	0.95 (d, δCH_3); 0.96 (d, $\delta' CH_3$); 1.71 (m, γCH)/26.56; 1.72 (m, βCH_2)/ 39.94; 3.74 (t, αCH)/56.10
25	Lysine	1.47 (m, γCH_2); 1.71 (m, δCH_2); 1.89 (m, βCH_2); 3.03 (t, εCH_2); 3.76 (t, αCH)
26	Methanol	3.35 (s, CH_3)
27	Methionine	2.13 (s, S-CH_3); 2.14 (m, βCH); 2.21 (m, β'CH); 2.64 (t, γCH_2)
28	Phenylalanine	3.10 (dd, βCH); 3.29 (dd, β'CH); 3.97 (dd, αCH); 7.32 (m, C2H, C6H, ring); 7.37 (m, C4H, ring); 7.42 (C3H, C5H, ring)
29	Phospholipids (mainly in HDL):	
	- fatty acyl chains	0.84 (br,CH_3)/14.10; 1.23 (br, $(CH_2)_n$)/31.75;1.58 (br, $\underline{CH_2}$-CH_2-CO); 2.00 (br, CH=CH-$\underline{CH_2}$)/27.11; 2.22 (br, CH_2-CO)/33.75; 2.68/2.74 (br, CH=CH-$\underline{CH_2}$-CH=CH)/25.60; 5.27 (br, CH=CH)/ 127.88;
	- choline headgroup	3.21 (br s, $N(CH_3)_3$)/54.16; 3.68 (br, CH_2(NH))/66.12; 4.30 (br, CH_2(OH))/59.60
30	Pyruvate	2.36 (s, CH_3)
31	Threonine	1.31 (d, γCH_3); 3.58 (d, αCH); 4.29 (m, βCH)
32	Triglycerides (mainly in LDL+VLDL)	
	- fatty acyl chains	0.87 (br, CH_3); 1.27 (br, $(CH_2)_n$)/29.57; 1.58 (br, $\underline{CH_2}$-CH_2-CO); 2.00 (br, CH=CH-$\underline{CH_2}$)/27.11; 2.22 (br, CH_2-CO)/33.75; 2.68/2.74 (br, CH=CH-$\underline{CH_2}$-CH=CH)/25.60; 5.29 (br, CH=CH)/129.53
	- glyceryl backbone	4.05, 4.28 (br, $C1H_2/C3H_2$); 5.20 (br, C2H)
33	Tyrosine	6.89 (d, C3H, C5H, ring); 7.18 (d, C2H, C6H, ring)
34	Valine	0.98 (d, γCH_3); 1.03 (d, $\gamma' CH_3$); 2.25(m, βCH); 3.62 (d, αCH)/66.18

4.2 Potential of plasma NMR profile to discriminate between patients and control subjects

Blood is a major vehicle of gases, nutrients, hormones and waste products in the body, mediating the metabolic interactions between different organs/tissues. It is therefore

reasonable to expect that the metabolic content of blood plasma may mirror the biochemical events accompanying a disease such as lung cancer. A first look at the plasma average spectra from control and cancer subjects (Figure 4.3) suggested differences in the lipoprotein profile and variations in the levels of some small metabolites, namely lactate, valine, glutamine, methanol and histidine.

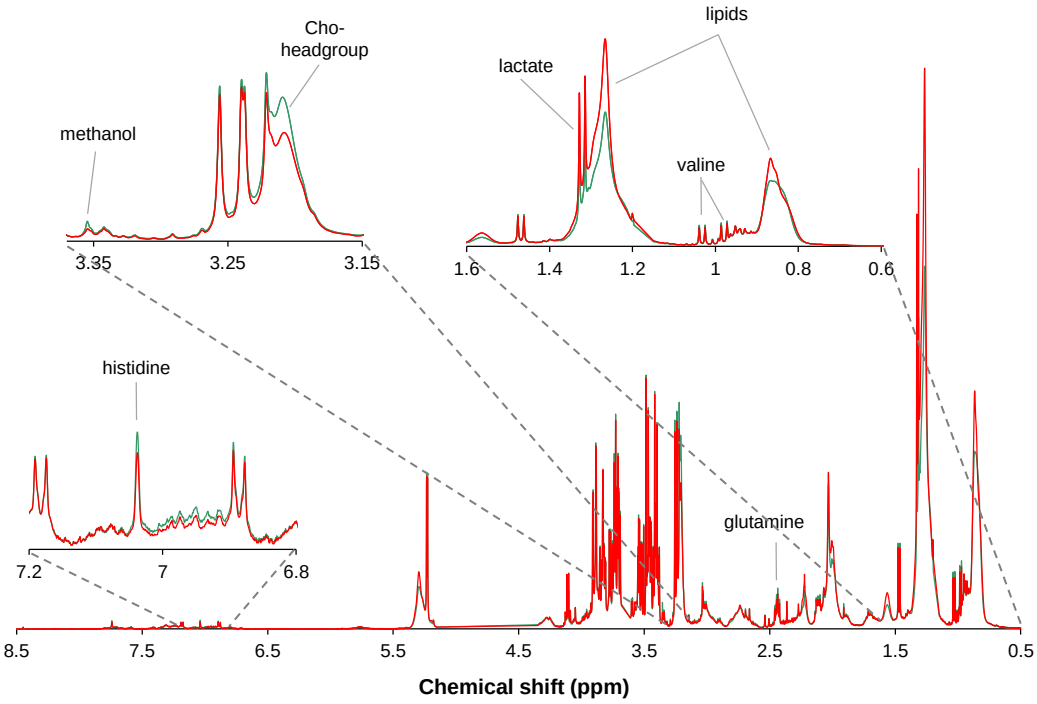


Figure 4.3 500 MHz average CPMG ^{1}H NMR spectra of blood plasma from controls (n 94, green) and lung cancer patients (n 106, red), after PQN normalisation.

Multivariate analysis was then applied to verify the consistency of these and other variations. The first step consisted of applying PCA, PLS-DA and OPLS-DA to the three NMR experiments recorded (standard 1D, CPMG and diffusion-edited), followed by Monte-Carlo cross validation (MCCV) and permutation testing of PLS-DA models. Although all three models performed similarly (Table 4.2), the model built with CPMG spectra showed slightly higher median Q^2 (0.59) and sensitivity (85.6%), resulting in an overall classification rate of 86.6%. In addition, as the broad background of proteins overlapping small metabolites was highly reduced in CPMG, the corresponding loadings were more informative concerning small metabolites' variations. Thus, the CPMG experiment was selected for spectral and multivariate analysis.

Table 4.2 Prediction results obtained by MCCV (500 iterations) of plasma PLS-DA models assessing the discrimination between lung cancer patients and healthy controls.

PLS-DA models	Median Q^2	Sensitivity (%)	Specificity (%)	Classification rate (%)
Standard 1D	0.53	79.3	91.6	85.1
CPMG	0.59	85.6	87.8	86.6
Diffusion-edited	0.54	83.0	92.4	87.4
12 integrals	0.56	87.9	85.2	86.6

As it can be seen in Figure 4.4a, control and cancer samples largely overlap in the PCA scores scatter plot, their separation being improved in PLS-DA and OPLS-DA score scatter plots (Figure 4.4b and 4.4c, respectively). MCCV and permutation testing were then used to verify model robustness. The classification results plotted in the ROC space (Figure 4.4d) showed that the majority of the real models (iterations with true classes assigned) afforded high sensitivity and specificity, whereas the permuted models (class membership ramdomly attributed) fell along the diagonal, i.e., the line of no discrimination (classification rate about 50%). Moreover, the Q^2 values were high when the true classes were assigned and negative or very low for most permuted models (Figure 4.4e), thus validating the predictive ability of the real model.

The identification of the metabolite variations responsible for class discrimination was then performed by inspection of OPLS-DA LV1 loadings, coloured according to variable importance in the projection (VIP), as depicted in Figure 4.4f. Lactate, acetoacetate and pyruvate showed positive loadings, indicating their increased levels in lung cancer patients, whereas methanol and several amino acids (e.g., valine, arginine/lysine, glutamine, serine and histidine) presented negative loadings, thus lower levels in patients compared to controls. Moreover, two other unidentified small resonances in the aromatic region seemed to be important for class separation, but their identification was not possible, as they did not show relevant correlations, either in 2D spectra or in STOCSY plots.

Regarding lipoprotein signals, some showed positive loadings, while others had negative loadings, suggesting a different proportion of the main lipoprotein subclasses in the plasma samples from control and cancer subjects. In particular, negative loadings (decreased in cancer patients) were found for broad resonances at δ 0.83, 1.23, 3.21 and 5.26, assigned to phospholipids composing mainly HDL lipoproteins.

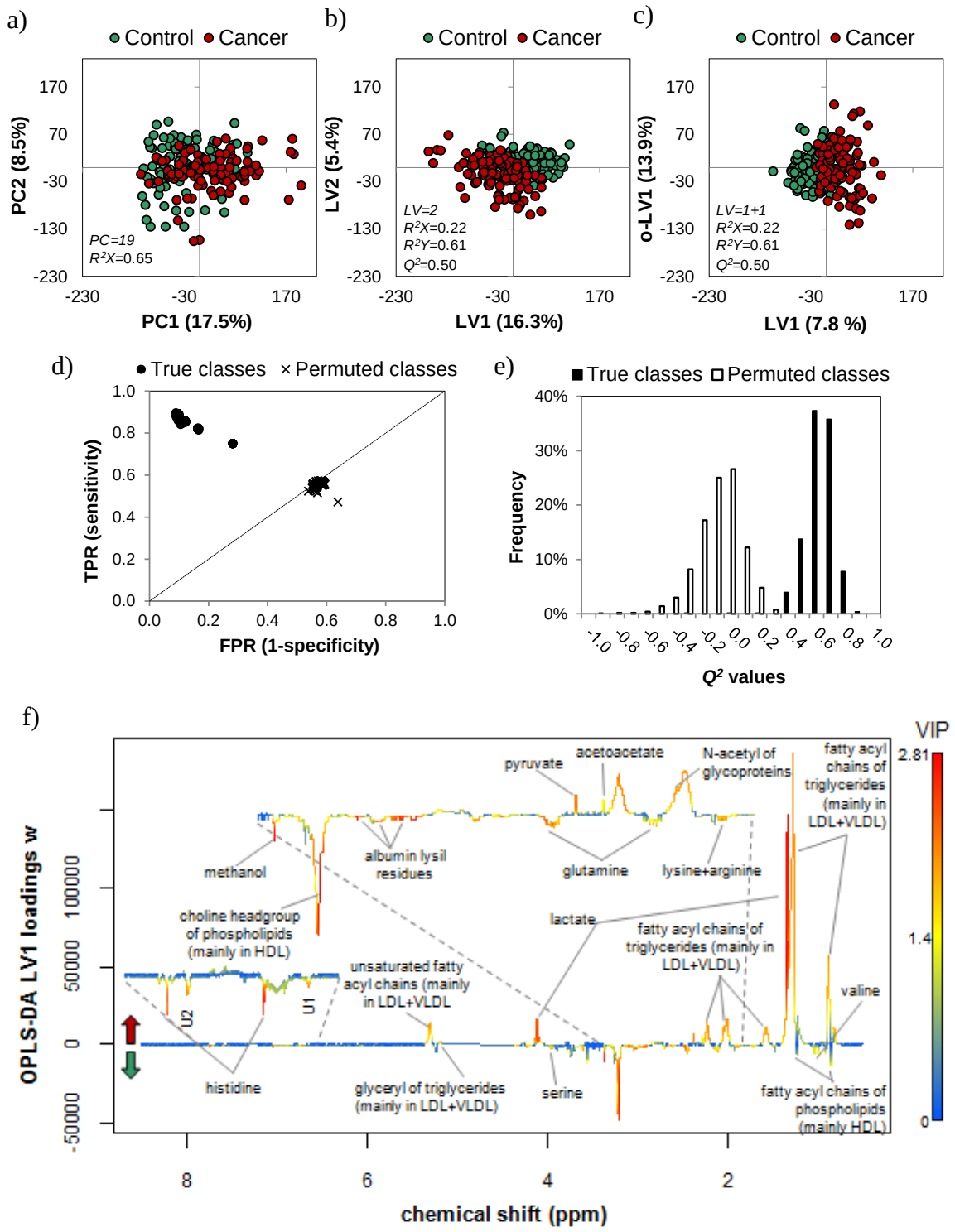


Figure 4.4 MVA applied to the CPMG ^{1}H NMR spectra of blood plasma from controls (n 94) and cancer patients (n 106): a) PCA, b) PLS-DA and c) OPLS-DA scores scatter plot. The parameters shown on the scores plots (*PC*: principal components, *LV*: latent variable, R^2X: variation explained by the X matrix, R^2Y: variation explained by the Y matrix, Q^2: predictive power) derive from default 7-fold cross validation. d) ROC space (TPR: true positive rate, FPR: false positive rate) and e) Q2 histogram obtained by MCCV and permutation testing (500 iterations) of the PLS-DA model. f) OPLS-DA LV1 loadings weights coloured as a function of VIP. Metabolites showing VIP>1 are assigned in the plot (U1: unkown 1, δ 6.71; U2: unknown 2, δ 2.72).

On the other hand, the signals assigned to triglycerides composing LDL+VLDL lipoproteins showed positive loadings, indicating their increased levels in patients. Other variations observed were the negative loadings for the albumin resonances, indicating a decrease of this protein levels in cancer patients, whereas glycoproteins appeared to be increased (positive loadings for their *N*-acetylated groups).

Following the thorough inspection of the loadings plot, metabolites with VIP>1 were integrated in the CPMG spectra, the median values of each group statistically compared by the Wilcoxon rank sum test and the corresponding effect size (*d*) calculated. Some resonances, namely those of albumin, lysine, arginine and serine, could not be integrated due to severe spectral overlap. Table 4.3 summarises the integration results for the metabolites showing statistically significant differences ($p<4.2\times10^{-3}$, Bonferroni-corrected) between cancer patients and controls. Confirming the observations derived from the loadings inspection, acetoacetate, lactate, pyruvate, LDL+VLDL lipoproteins and glycoproteins were increased in patients' plasma, whereas HDL lipoproteins, amino acids (glutamine, histidine and valine), methanol and two unknown peaks in the aromatic region (δ 6.71 and δ 7.59) were found to be significantly decreased compared to controls.

Table 4.3 Blood plasma metabolites showing statistically significant differences between lung cancer patients and healthy controls. For each metabolite, the average percentage and coefficient of variation were obtained by spectral integration of selected signals. Effect sizes and *p*-values are shown, indicating, respectively, the magnitude and statistical significance of the differences between the two groups. n.s. not significant; ↓ qualitative decreased (signal not integrated due to spectral overlap); s: singlet; d: doublet; t: triplet; q: quartet; m: multiplet; br: broad; br s: broad singlet.

Metabolite (δ, multiplicity)	**Cancer vs. Control**		
	% variation	***p*-value[a]**	**effect size**
Acetoacetate (2.27, s)	19.6±4.0	5.1×10^{-7}	0.62±0.28
Albumin (3.00, br)		↓	
Arginine + lysine (1.92, m)		↓	
Glutamine (2.44, m)	-14.0±2.2	3.8×10^{-12}	-1.02±0.30
Glycoproteins (2.02+2.05, br s)	11.7±1.1	$<2.2\times10^{-16}$	1.40±0.31
Histidine (7.04, s)	-17.7±2.3	2.5×10^{-16}	-1.31±0.31
Lactate (4.11, q)	32.6±3.5	4.9×10^{-14}	1.10±0.30
Methanol (3.35, s)	-33.0±7.5	2.6×10^{-9}	-0.83±0.29
HDL (3.21, br)	-19.0±2.5	2.6×10^{-14}	-1.25±0.30
Pyruvate (2.36, s)	22.2±2.8	6.0×10^{-12}	0.99±0.29
Serine (3.96, m)		↓	

Table 4.3 (continued)

Metabolite (δ, multiplicity)	Cancer vs. Control		
	% variation	*p*-value[a]	effect size
LDL+VLDL (0.87, br)	31.8±3.7	2.7×10^{-11}	1.02±0.30
Valine (1.03, d)	-8.5±2.2	5.0×10^{-5}	-0.90±0.28
U1 (6.71, br)	-31.7±5.4	1.3×10^{-13}	-1.08±0.30
U2 (7.59, br)	-26.1±3.7	5.2×10^{-14}	-1.19±0.30

[a]Wilcoxon rank sum test $p<4.2\times10^{-3}$ (Bonferroni-corrected).

The classification ability of the twelve metabolites integrated was then evaluated by applying PLS-DA to their signal areas. As shown in Table 4.2 and Figure 4.5, the MCCV results obtained for this model were quite similar to those obtained when modelling the full resolution spectra, thus confirming the relevance of these metabolic features as possible cancer-related markers.

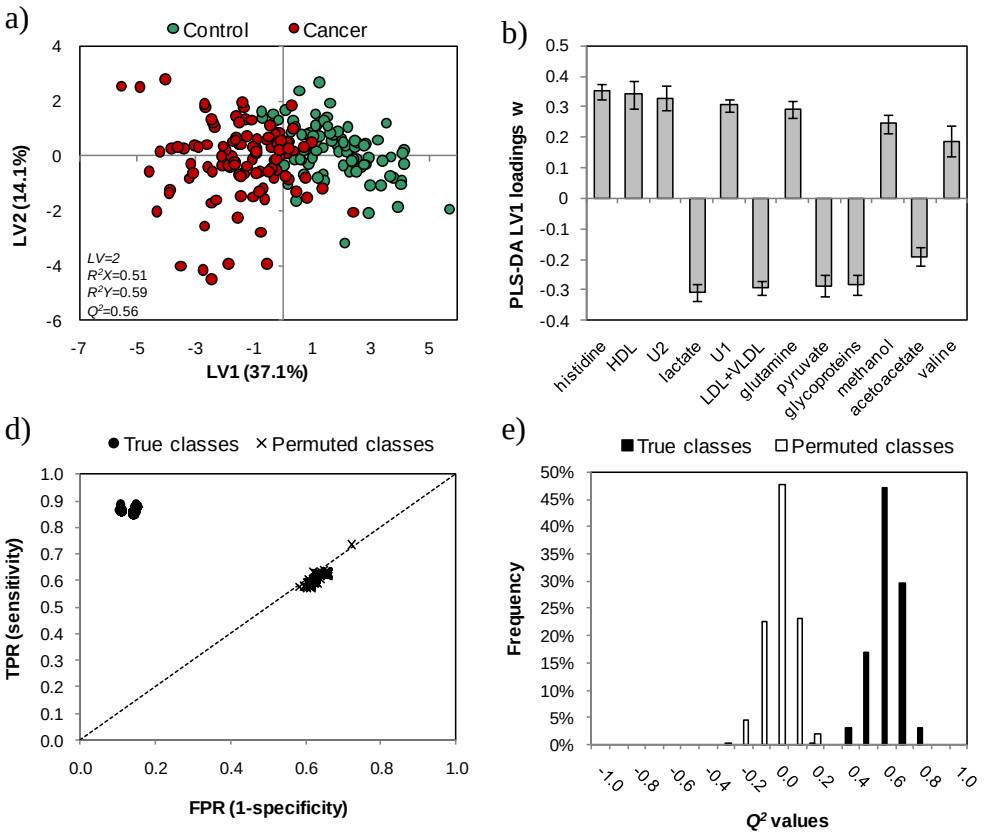


Figure 4.5 PLS-DA applied to 12 signal integrals measured in the CPMG ^{1}H NMR spectra of blood plasma from controls (n 94) and cancer patients (n 106): a) scores scatter plot and b) LV1 loadings weights. The parameters shown on the scores plot (*LV*: latent variable, R^2X: variation explained by the X matrix, R^2Y: variation explained by the Y matrix, Q^2: predictive power) derive from default 7-fold cross validation. c) ROC space (TPR: true positive rate, FPR: false positive rate) and d) Q^2 histogram obtained by MCCV and permutation testing (500 iterations).7

4.3 Impact of tumour histological type on the plasma metabolic composition

As patients included in this study were diagnosed to have tumours of different histological type, the possible impact of such heterogeneity on blood plasma composition has also been investigated. Firstly, the discrimination between the control group and each of the two subgroups corresponding to the main histological types (adenocarcinoma – AdC, and squamous cell carcinoma – SqCC) was assessed by PLS-DA and MCCV. As presented in Table 4.4, the predictive power for discriminating between controls and AdC (Q^2 0.50) or between controls and SqCC (Q^2 0.46) was lower than the one obtained for the model comprising all samples (Q^2 0.59). Also, lower sensitivity, specificity and classification rate were obtained, although this might be partially due to the unbalanced number of samples between the cancer and control groups, as using a smaller control group (n 40) afforded an improvement in classification parameters (Table 4.4).

Table 4.4. Prediction results obtained by MCCV (500 iterations) of plasma PLS-DA models, assessing the discrimination between different groups of samples (including different histological types and stages). AdC: adenocarcinoma; SqCC: squamous cell carcinoma.

PLS-DA classes [**n samples**; n M/n F; median age]	**Median Q^2**	**Sensitivity (%)**	**Specificity (%)**	**Classification rate (%)**
All samples: Control vs. Cancer [**94**; 49/45; 44] vs. [**106**; 76/30; 63]	0.59	85.6	87.8	86.6
Control vs. AdC [**94**; 49/45; 44] vs. [**40**; 23/17; 66]	0.50	68.2	67.6	67.9
Control vs. SqCC [**94**; 49/45; 44] vs. [**27**; 25/2; 63]	0.46	64.0	80.5	73.9
Control vs. AdC[a] [**40**; 22/18; 48] vs. [**40**; 23/17; 66]	0.57	73.5	95.9	89.2
Control vs. SqCC[a] [**40**; 22/18; 48] vs. [**27**; 25/2; 63]	0.40	66.3	97.1	90.2
AdC vs. SqCC [**40**; 23/17; 66] vs. [**27**; 25/2; 63]	-0.39	21.9	63.3	46.6
AdC vs. SqCC[b] [**40**; 23/17; 66] vs. [**27**; 25/2; 63]	0.03	50.3	74.1	64.5
Control vs. Cancer stage I [**94**; 49/45; 44] vs. [**61**; 44/17; 63]	0.60	82.0	93.9	89.2
Cancer stage I vs. Cancer stage II+III [**61**; 44/17; 63] vs. [**34**; 25/9; 62]	-0.19	22.2	74.2	55.6
Cancer stage I vs. Cancer stage II+III[b] [**61**; 44/17; 63] vs. [**34**; 25/9; 62]	0.41	77.2	87.1	83.5

[a] Using balanced sample numbers in the groups compared. [b]After variable selection.

In order to assess if AdC and SqCC tumours impact a different metabolic signature on blood plasma of cancer patients, the loadings of the corresponding OPLS-DA models (not shown) were analysed and compared to the loadings of the model based on the complete dataset. The variations between controls and each histological type were then calculated for signals with VIP>1 and are listed in Table 4.5. Interestingly, although most significant alterations were common to both histological types (showing similar % variations in relation to controls), others seemed to be specific for each tumour type. In particular, alanine was found to be significantly decreased in SqCC but not in AdC (or for the whole set of patients), while β-hydroxybutyrate was increased in the plasma of AdC patients only. Furthermore, the decrease in valine levels was not significant in SqCC, and acetate (VIP<1 when considering all samples) was found to be important and significantly decreased when taking into account each histological type separately (Table 4.5).

In spite of the few differences described above, the potential discrimination between AdC and SqCC patients was attempted by PLS-DA of their plasma spectra (leaving out the controls). However, none of the models built (based on full resolution data or after the variable selection procedure described in subchapter 2.3.5 could be MCCV-validated, showing negative or near zero Q^2 and poor classification parameters (Table 4.4). Thus, while showing great promise in cancer vs. control discrimination, plasma NMR metabolomics failed to distinguish between AdC and SqCC patients.

4.4 Impact of tumour stage on the plasma metabolic composition

In order to assess the possible influence of disease stage on the plasma profiles, the patients group was divided into subsets of stages I (n 61), II (n 24) and III (n 10) and multivariate analysis was performed to evaluate the differences between each of these subsets and the control group (n 94). PLS-DA of stage I patients vs. controls produced an MCCV-validated model with a predictive power and classification rate comparable to those obtained for the whole dataset (Table 4.4, Figure 4.6a, b). However, there was a slight decrease in sensitivity that could be attributable to the lower impact of early stage tumours on blood plasma composition, and/or, to some extent, to the unbalanced number of samples in each group. In fact, when a model was built with the same number of samples in both groups (controls and stage I), sensitivity improved to 87% (not shown).

Table 4.5 Blood plasma metabolites showing statistically significant differences between lung cancer patients and healthy controls, considering all samples in the patients group (2nd column), only adenocarcinomas (3rd column) or only squamous cell carcinomas (4th column). For each metabolite, the average percentage and coefficient of variation were obtained by spectral integration of selected signals. Effect sizes and *p*-values are shown, indicating, respectively, the magnitude and statistical significance of the same differences. n.s. not significant; ↓ qualitative decreased (signal not integrated due to spectral overlap); s: singlet; d: doublet; t: triplet; q: quartet; m: multiplet; br: broad; br s: broad singlet.

Metabolite (δ, multiplicity)	**Cancer vs. Control**			**Cancer vs. Control (AdC)**			**Cancer vs. Control (SqCC)**		
	% variation	**effect size**	***p*-value[a]**	**% variation**	**effect size**	***p*-value[a]**	**% variation**	**effect size**	***p*-value[a]**
Acetate(1.92, s)		VIP<1		-8.3±2.9	-0.54±0.38	1.2×10^{-3}	-10.8±3.1	-0.71±0.44	1.1×10^{-3}
Acetoacetate (2.27, s)	19.6±4.0	0.62±0.28	5.1×10^{-7}	20.7±6.9	0.70±0.38	1.4×10^{-3}	22.5±7.9	0.85±0.44	1.6×10^{-4}
Alanine (1.48, d)		n.s.			n.s.		-10.1±2.7	-0.81±0.44	1.4×10^{-3}
Albumin (3.00, br)		↓			↓			↓	
Arginine + lysine (1.92, m)		↓			↓			↓	
β-Hydroxybutyrate (2.31, m)		n.s.		17.2±3.9	0.83±0.38	8.0×10^{-4}		n.s.	
Glutamine (2.44, m)	-14.0±2.2	-1.02±0.30	3.8×10^{-12}	-12.8±2.7	-0.99±0.39	3.0×10^{-7}	-15.3±3.6	-1.15±0.45	1.9×10^{-5}
Glycoproteins (2.02+2.05, br s)	10.5±1.4	0.98±0.29	3.5×10^{-10}	7.9±1.9	0.88±0.38	4.6×10^{-5}	13.5±2.7	1.42±0.46	1.2×10^{-6}
HDL (3.21, br)	-19.0±2.5	-1.25±0.30	2.6×10^{-14}	-18.1±2.8	-1.22±0.40	1.5×10^{-8}	-24.8±3.5	-1.63±0.47	2.1×10^{-9}
Histidine (7.04, s)	-17.7±2.3	-1.31±0.31	2.5×10^{-16}	-18.1±2.8	-1.29±0.40	2.5×10^{-10}	-18.3±3.4	-1.28±0.46	1.6×10^{-6}
Lactate (4.11, q)	32.6±3.5	1.10±0.30	4.9×10^{-14}	32.1±5.4	1.25±0.40	1.1×10^{-7}	31.1±7.1	1.22±0.45	2.3×10^{-6}
LDL+VLDL (0.87, br)	31.8±3.7	1.02±0.30	2.7×10^{-11}	24.9±5.2	0.96±0.39	7.0×10^{-6}	35.5±7.1	1.32±0.46	2.1×10^{-6}
Methanol (3.35, s)	-33.0±7.5	-0.83±0.29	2.6×10^{-9}	-30.3±8.0	-0.72±0.38	3.9×10^{-5}	-39.0±8.3	-0.94±0.44	8.4×10^{-9}
Pyruvate (2.36, s)	22.2±2.8	0.99±0.29	6.0×10^{-12}	22.2±3.2	1.29±0.40	1.5×10^{-9}	18.9±3.7	1.11±0.45	3.2×10^{-6}
Serine (3.96, m)		↓			↓			↓	
Valine (1.03, d)	-8.5±2.2	-0.61±0.28	5.0×10^{-5}	-9.7±2.8	-0.64±0.38	5.8×10^{-4}		n.s.	
U1 (6.71, br)	-31.7±5.4	-1.08±0.30	1.3×10^{-13}	-36.3±5.7	-1.20±0.40	2.4×10^{-10}	-29.3±6.6	-0.94±0.44	1.7×10^{-5}
U2 (7.59, br)	-26.1±3.7	-1.19±0.30	5.2×10^{-14}	-24.2±4.1	-1.24±0.40	1.4×10^{-8}	-29.8±5.6	-1.45±0.47	2.6×10^{-7}

[a] Wilcoxon rank sum test $p<4.2\times10^{-3}$ (Bonferroni-corrected).

Inspection of the corresponding loadings then showed that the metabolic differences discriminating controls and stage I patients were highly coincident with those identified when considering all three stages together. This is clearly visualized through a VIP-wheel representation (devised in-house by colleagues, Diaz et al. 2013), which shows great superimposition of the important variables (VIP>1) highlighted in the two models (whole dataset and control vs. stage I) (Figure 4.6c). Therefore, this result demonstrates that a putative cancer signature is detectable in plasma right from early phases of tumour development. Also, it allows excluding the possible metabolic impact of malnutrition and weight loss, often affecting advanced stage patients (Okamoto et al. 2009), as an important confounder in the whole dataset results. The same kind of analysis was pursued for stage II and stage III, and, although PLS-DA models could not be properly validated (likely due to the unbalanced number of samples in control and cancer groups), the metabolic variations highlighted were again very similar to those identified when considering all samples.

Additionally, the pairwise comparison of all three tumour stages was attempted: stage I vs. stage II, stage I vs. stage III and stage I vs. stage II+III. In all cases no valid discrimination was achieved (Q^2<0 and sensitivities below 30%), so that a method of variable selection was applied (as described in section 2.3.5). The best result was achieved for the discrimination between stage I and stage II+III, for which MCCV afforded a classification rate of 83% (Table 4.4). When analysing the variables selected, a few metabolites were suggested to be responsible for this separation, namely acetate, glutamine, formate and tyrosine (increased in stage I samples), and acetoacetate, pyruvate, β-hydroxybutyrate and lactate (increased in stage II+III samples). However, none of these variations were confirmed to be statistically significant by spectral integration, thus hindering solid conclusions.

4.5 Influence of potential confounders in plasma-based cancer vs. control discrimination

Blood plasma metabolic content is considered to be relatively stable, accounting for its homeostatic role in the body. However, several extrinsic and intrinsic factors (e.g., diet, lifestyle, age, gender) may be responsible for some inter-individual variability, and influence the interpretation of disease-related effects. For that reason, the groups

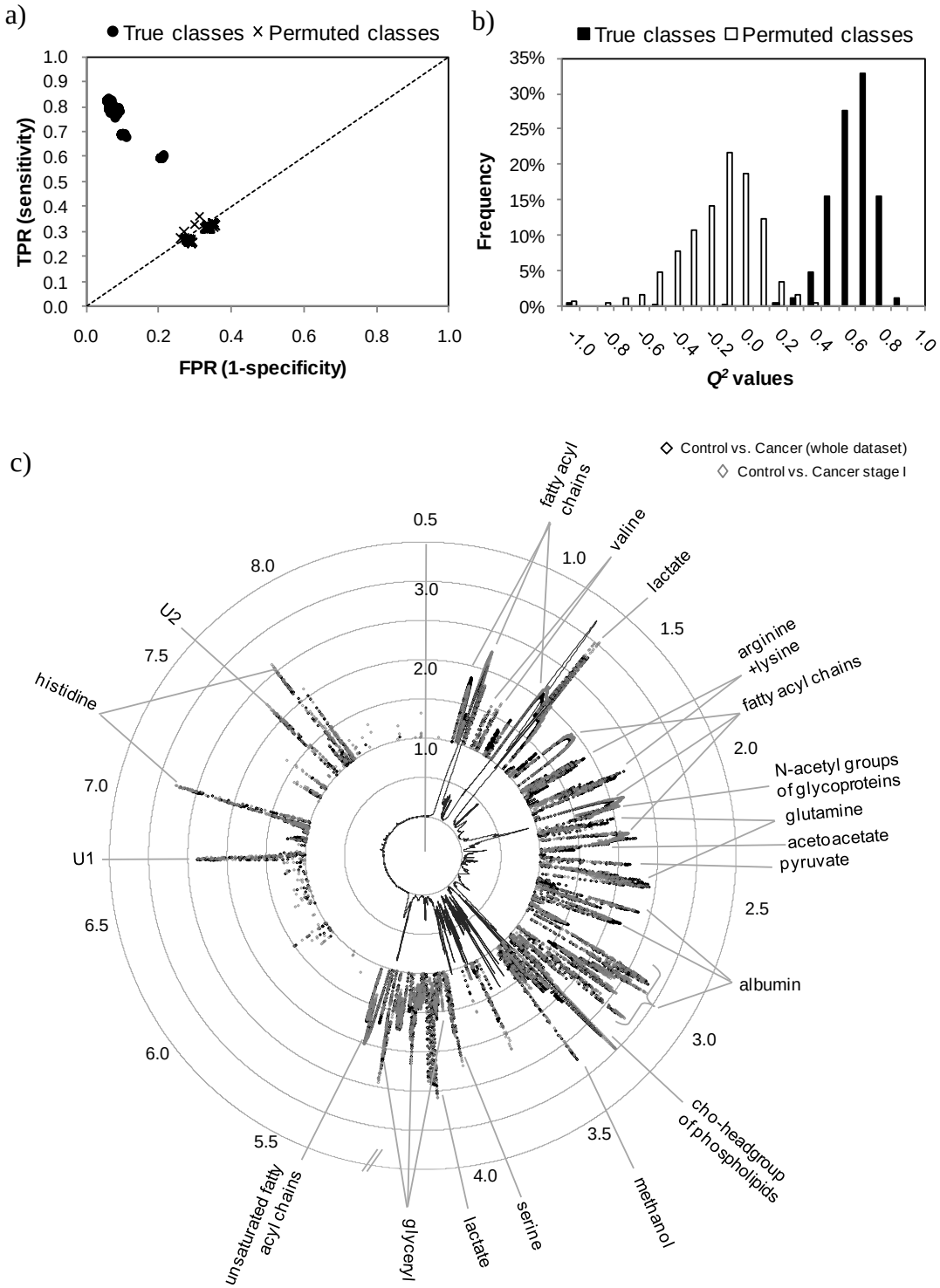


Figure 4.6 a) ROC space (TPR: true positive rate, FPR: false positive rate) and b) Q^2 histogram obtained by MCCV and permutation testing (500 iterations) of the PLS-DA model built with ^{1}H NMR spectra of blood plasma to assess control vs. cancer stage I discrimination. c) VIP-wheel representation of variables found to have VIP>1 in OPLS-DA models built with: the whole dataset (black), the subset with controls and stage I patients (grey).

compared (e.g., patients vs. healthy controls) should ideally be matched in terms of possible confounding factors. In practice, this condition is often difficult to fulfil, so that the contribution of those factors to metabolic variability should be assessed. In this work, the influence of gender, age and smoking habits on the plasma metabolic profile and on the disease-related classification models was investigated.

4.5.1 Gender-related metabolic features in blood plasma

Regarding gender, the control group was well balanced with ca. 50% of males and females, whilst the cancer group included a higher number of male subjects (72%) (p<0.01), and thus possible gender-related bias was evaluated. In a first approach, the effect of gender on the metabolic profile of blood plasma was assessed by considering only the control group. The MCCV results of the corresponding PLS-DA model (male vs. female) afforded a good median predictive ability (Q^2 0.53), and a classification rate of 87% (Table 4.6).

Table 4.6 Prediction results obtained by MCCV (500 iterations) of plasma PLS-DA models assessing the discrimination between different groups of samples (including groups varying in gender, age and smoking habits).

PLS-DA classes [**n samples**; n M/n F; median age]	**Median Q^2**	**Sensitivity (%)**	**Specificity (%)**	**Classification rate (%)**
Control vs. Cancer [**94**; 49/45; 44] vs. [**106**; 76/30; 63]	0.59	85.6	87.8	86.6
Male vs. Female, controls only [**49**; 49/0; 44] vs. [**45**; 0/45; 45]	0.53	84.9	89.0	87.0
Control vs. Cancer, males only [**49**; 49/0; 44] vs. [**76**; 76/0; 64]	0.58	91.4	80.1	87.0
Control vs. Cancer, females only [**45**; 0/45; 45] vs. [**30**; 0/30; 58]	0.47	71.3	88.1	81.4
Control vs. Cancer, age 41-60 [**42**; 21/21; 50] vs. [**38**; 24/14; 56]	0.52	66.8	71.9	69.4
Smoker vs. Never-smoker, controls only [**29**; 17/12; 38] vs. [**44**; 19/25; 45]	-0.24	46.0	66.2	58.2
Control smoker vs. Cancer[a] [**29**; 17/12; 38] vs. [**45**; 33/12; 60]	0.54	84.9	85.9	85.3
Control never-smoker vs. Cancer[a] [**44**; 19/25; 45] vs. [**45**; 33/12; 60]	0.58	83.0	88.1	85.5

[a] Using balanced sample numbers in the groups compared.

The separation between genders was clear in the OPLS-DA scores scatter plot (Figure 4.7a) and the metabolic features explaining this discrimination were revealed by the corresponding loadings (Figure 4.7b). The main differences between healthy males and females were in the lipoprotein profile, with HDL increased in females and

LDL+VLDL increased in males, in the levels of creatine (increased in females) and of creatinine and valine (increased in males). These findings are in agreement with other studies reporting the same gender-related variations in the plasma of healthy subjects (Kochhar et al. 2006; Lawton et al. 2008).

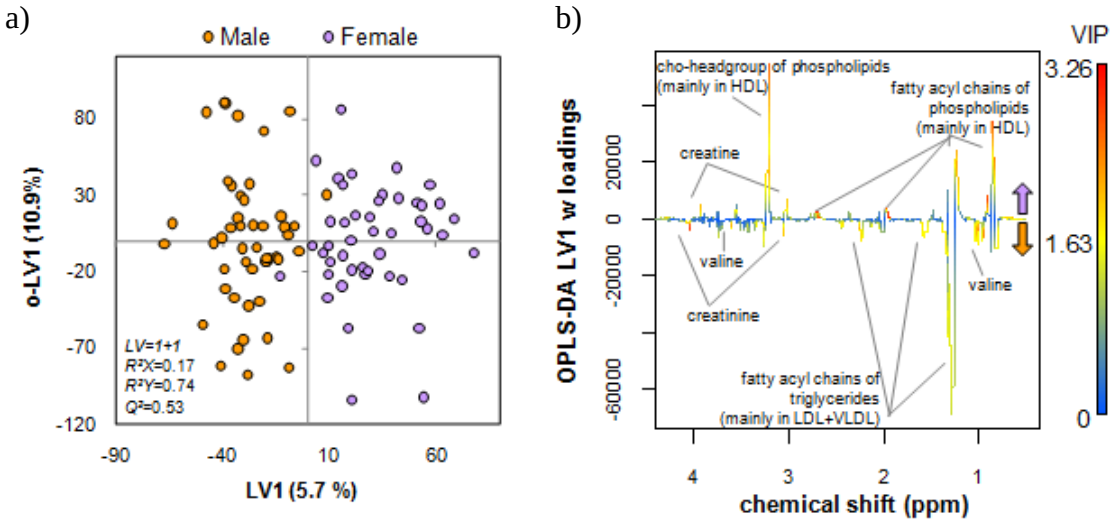


Figure 4.7 a) OPLS-DA scores scatter plot (*LV* 1+1, R^2X 0.17, R^2Y 0.74, Q^2 0.53) of the CPMG spectra of males (n 49) and females (n 45) of the control group, and b) corresponding loadings weights, coloured according to VIP parameter. The parameters shown on the scores plot (*LV*: latent variable, R^2X: variation explained by the X matrix, R^2Y: variation explained by the Y matrix, Q^2: predictive power) derive from default 7-fold cross validation.

By confronting these variations with putative disease-related effects (Table 4.7), the metabolic features showing the same variations in males (vs. females) and in the cancer group (where males predominated) were highlighted as possibly biased. This was the case of HDL and LDL+VLDL lipoproteins, which thus required further attention (as described below).

Table 4.7 Gender-related metabolic variations in plasma and their comparison with cancer-related variations

Gender-related metabolites	Variation male vs. female	Variation cancer vs. control	Observations
Creatine	↓	-	
Creatinine	↑	-	
HDL	↓	↓	Possible bias[a]
LDL+VLDL	↑	↑	Possible bias[a]
Valine	↑	↓	Opposite variation

[a] Metabolites increased/decreased in both male and cancer groups may potentially bias cancer vs. control discrimination as the cancer group comprises ~70% male subjects.

Another approach to assess the influence of gender on cancer vs. control discrimination consisted of building separate models for males and females and reassessing the cancer-related variables. The two models showed good prediction parameters, although the female model showed decreased predictive power (Q^2 0.47) and classification rate (81%) compared to the male and the whole dataset models (Table 4.6). In regard to discriminant metabolites, there was a large coincidence between the variables with VIP>1 resulting from the different OPLS-DA models compared (Figure 4.8).

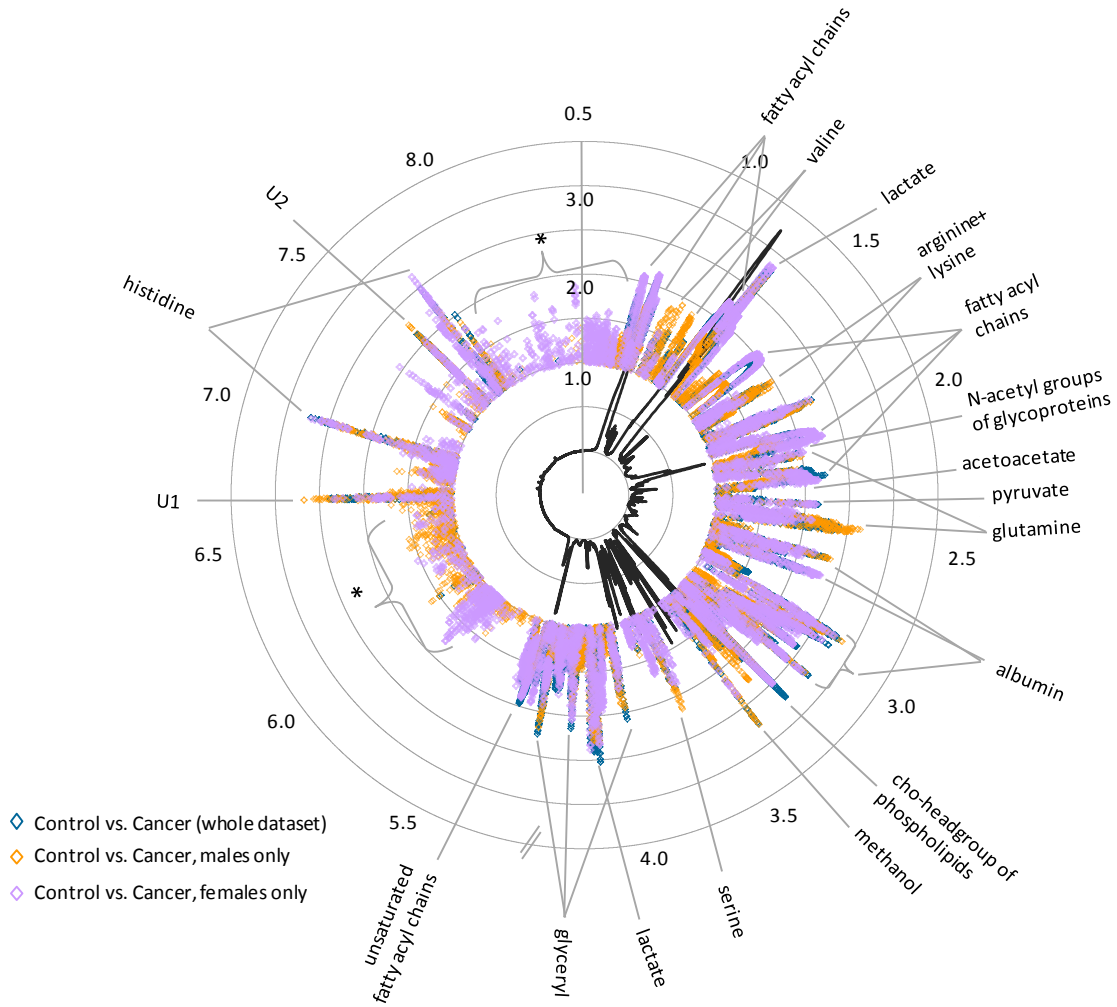


Figure 4.8 VIP-wheel representation of variables found to have VIP>1 in the plasma OPLS-DA models built with: the whole dataset (blue), male subjects only (orange) and female subjects only (purple).

Table 4.8 Blood plasma metabolites showing statistically significant differences between lung cancer patients and healthy controls, considering all samples in the patients group (2nd column), only males (3rd column), only females (4th column) or an age-matched subset (5th). For each metabolite, the average percentage and coefficient of variation were obtained by spectral integration of selected signals. Effect sizes and *p*-values are shown, indicating, respectively, the magnitude and statistical significance of the differences between the two groups. n.s. not significant; ↓ qualitative decreased (signal not integrated due to spectral overlap); s: singlet; d: doublet; t: triplet; q: quartet; m: multiplet; br: broad; br s: broad singlet.

Metabolite (δ, multiplicity)	**Cancer vs. Control (ALL SAMPLES)**			**Cancer vs. Control (MALES ONLY)**			**Cancer vs. Control (FEMALES ONLY)**			**Cancer vs. Control (AGE-MATCHED)**		
	% variation	***p*-value[a]**	**effect size**	**% variation**	***p*-value[a]**	**effect size**	**% variation**	***p*-value[a]**	**effect size**	**% variation**	***p*-value[a]**	**effect size**
Acetate		VIP<1			VIP<1			VIP<1		-13.3±3.6	1.7×10^{-4}	-0.86±0.46
Acetoacetate (2.27, s)	19.6±4.0	5.1×10^{-7}	0.62±0.28	24.3±5.1	4.1×10^{-5}	0.67±0.37		n.s.		17.9±5.6	6.1×10^{-4}	0.69±0.45
Alanine (1.48, d)		n.s.			n.s.			n.s.		-9.5±3.0	1.8×10^{-3}	-0.75±0.45
Albumin (3.00, br)		↓			↓			↓			↓	
Arginine + lysine (1.92, m)		↓			↓			↓			↓	
Glucose (3.51, m)	-1.2±2.4	3.8×10^{-3}	-0.07±0.28		n.s.			n.s.		-7.4±2.5	2.9×10^{-4}	-0.69±0.45
Glutamine (2.44, m)	-14.0±2.2	3.8×10^{-12}	-1.02±0.30	-16.8±2.6	4.7×10^{-10}	-1.22±0.39	-10.3±3.3	4.1×10^{-4}	-0.80±0.48	-15.2±2.8	5.9×10^{-7}	-1.33±0.48
Glycoproteins (2.02+ 2.05, br s)	11.7±1.1	$<2.2\times10^{-16}$	1.40±0.31	12.4±1.4	1.2×10^{-10}	1.37±0.40	10.2±1.8	5.1×10^{-7}	1.40±0.52	12.6±1.8	3.9×10^{-8}	1.52±0.50
HDL (δ 3.21, br)	-19.0±2.5	2.6×10^{-14}	-1.25±0.30	-17.3±2.9	1.6×10^{-8}	-1.24±0.39	-15.3±3.4	2.6×10^{-5}	-1.2±0.50	-16.1±3.7	9.6×10^{-6}	-1.05±0.47
Histidine (7.04, s)	-17.7±2.3	2.5×10^{-16}	-1.31±0.31	-20.0±3.0	4.1×10^{-11}	-1.42±0.40	-15.1±2.9	2.7×10^{-7}	-1.29±0.51	-12.6±3.2	2.7×10^{-4}	-0.95±0.46
Lactate (4.11, q)	32.6±3.5	4.9×10^{-14}	1.10±0.30	36.0±4.4	1.5×10^{-9}	1.12±0.38	20.6±5.4	2.2×10^{-4}	0.90±0.49	20.7±6.3	7.9×10^{-4}	0.68±0.45
LDL+VLDL (0.87, br)	31.8±3.7	2.7×10^{-11}	1.02±0.30	30.1±4.8	8.7×10^{-6}	0.90±0.38	29.4±5.9	4.6×10^{-6}	1.19±0.50	24.7±5.5	1.5×10^{-4}	0.92±0.46
Methanol (3.35, s)	-33.0±7.5	2.6×10^{-9}	-0.83±0.29	-38.2±10.3	1.5×10^{-6}	-0.94±0.38	-21.7±9.5	2.7×10^{-3}	-0.58±0.47	-33.7±11.9	1.3×10^{-3}	-0.74±0.45
Pyruvate (2.36, s)	22.2±2.8	6.0×10^{-12}	0.99±0.29	24.5±3.5	1.5×10^{-8}	1.02±0.38	19.2±4.8	6.8×10^{-5}	0.97±0.49	20.7±4.7	5.5×10^{-5}	0.91±0.46
Serine (3.96, m)		↓			↓			↓			↓	
Valine (1.03, d)	-8.5±2.2	5.0×10^{-5}	-0.90±0.28	-12.4±2.9	1.1×10^{-5}	-0.91±0.38		n.s.		-10.8±3.2	1.3×10^{-3}	-0.77±0.45
U1 (6.71, br)	-31.7±5.4	1.3×10^{-13}	-1.08±0.30	-37.6±7.2	3.8×10^{-11}	-1.34±0.40	-23.9±7.4	1.6×10^{-4}	-0.86±0.48	-24.1±7.1	1.2×10^{-4}	-0.85±0.46
U2 (7.59, br)	-26.1±3.7	5.2×10^{-14}	-1.19±0.30	-29.2±4.6	6.3×10^{-11}	-1.38±0.40	-21.3±6.4	3.1×10^{-4}	-0.94±0.49	-20.4±6.0	2.0×10^{-4}	-0.85±0.46

[a]Wilcoxon rank sum test $p<4.2\times10^{-3}$ (Bonferroni-corrected).

Indeed, spectral integration confirmed most differences between patients and controls to be independent of gender, since the % variation of most metabolites was similar when considering the two genders either together or separately (Table 4.8). This was valid for HDL and LDL+VLDL lipoproteins, which maintained their importance in individual gender models, thus allowing their possible gender-related bias to be discarded. On the other hand, variations in acetoacetate and valine were no longer significant in the discrimination between female controls and female patients (Table 4.8), suggesting that some metabolic alterations in cancer may be gender-specific.

4.5.2 Age-related metabolic features in blood plasma

This study had a major limitation with regard to age-matching of the control and cancer groups, as the average age of the controls was 42 years-old (median 44 years-old), while that of lung cancer patients was 62 years-old (median 63 years-old). This twenty-year difference ($p<2.2\times10^{-16}$) was due to the difficulty in finding older healthy volunteers to integrate the control group and to the greater prevalence of lung cancer in elderly subjects. To address the extent to which age could be reflected in the plasma metabolic profile, OPLS regression analysis was applied to the plasma spectra of control subjects using their age as classifier (age range 22-60). Although the resulting model showed a weak predictive ability (Q^2 near zero, as assessed by seven-fold cross validation), the distribution of scores along LV1 suggested some age-dependency (Figure 4.9a).

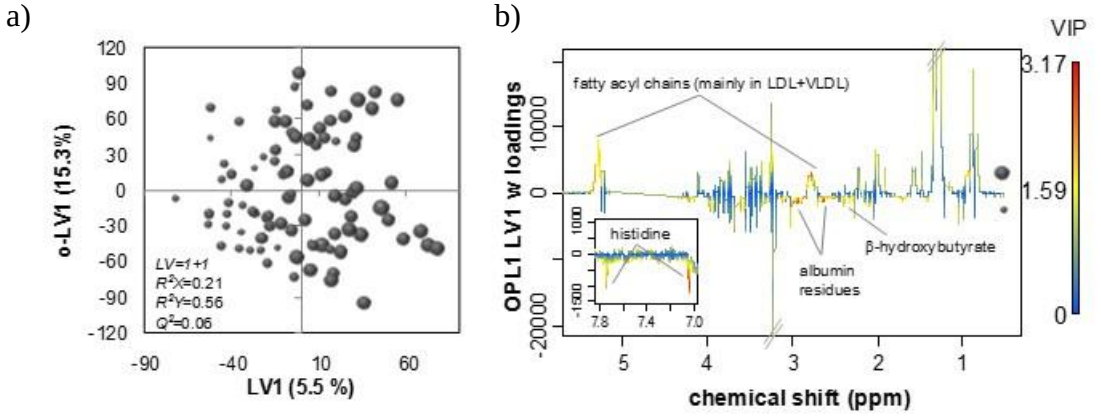


Figure 4.9 OPLS1 regression of the plasma ^{1}H NMR spectra of healthy controls (n 93) and the subjects age: a) scores scatter plot of the first two latent variables, where the size of the circle is proportional to age, b) LV1 loadings weights, coloured as a function of VIP. The parameters shown on the scores plot (*LV*: latent variable, R^2X: variation explained by the X matrix, R^2Y: variation explained by the Y matrix, Q^2: predictive power) derive from default 7-fold cross validation.

In the corresponding loadings (Figure 4.9b), the variables important for regression were the signals of lipoproteins LDL+VLDL (increased in older subjects), and β-hydroxybutyrate, histidine and albumin (increased in younger subjects). Concordantly, histidine has already been pointed out by Yu and colleagues as an age-dependent metabolite (Yu et al. 2012). The authors proposed the metabolic conversion of histidine to carnosine as a likely explanation for its lower levels in older people, since carnosine (a dipeptide formed of β-alanine and histidine) is considered to be a natural anti-aging molecule due to its role in suppressing biochemical alterations, like oxidative damage and protein glycation (Hipkiss 2010). Regarding lipoproteins, a study by Freedman and colleagues suggested positive correlations of VLDL and LDL concentrations with age, in particular for women, whilst HDL-age correlations were weak for both men and women (Freedman et al. 2004). As for albumin, an early paper reported that its concentration in human serum decreases with age in healthy subjects (Veering et al. 1990), thus agreeing with our results.

Taking into consideration that the cancer group was composed of older subjects, age could indeed be a source of bias regarding the variation of lipoproteins (increased in both cancer and older subjects), albumin and histidine (decreased in both cancer and older subjects) (Table 4.9), an hypothesis to be more deeply investigated, as described below.

Table 4.9 Age-related metabolic variations in plasma and their comparison with cancer-related variations.

Age-related metabolites	Variation with increasing age	Variation cancer vs. control	Observations
Albumin	↓	↓	Possible bias[a]
Histidine	↓	↓	Possible bias[a]
β-Hydroxybutyrate	↓	-	Opposite variation
LDL+VLDL	↑	↑	Possible bias[a]

[a] Metabolites increased/decreased in both older subjects and patients may potentially bias cancer vs. control discrimination as the median age of the cancer group is significantly higher than that of the control group.

In order to further assess possible influence of age on cancer vs. control discrimination a more age-balanced subset (n 80) (controls: n 42, 21M/21F, average/median age 52/50 years-old; cancer patients: n 38, 24M/14F, average/median age 55/56 years-old) has been tested. The resulting PLS-DA model could be MCCV-validated but showed much lower sensitivity and specificity than the whole dataset model,

affording an overall classification rate of 69% (Table 4.6). Thus, age mismatching could not be ruled out as an important confounding factor in the plasma-based cancer vs. control discrimination achieved in this study. In regard to discriminant metabolites, all variations found to be important in the whole dataset model held their importance in the subset with improved age-matching, as shown by the superimposition of variables in the VIP-wheel (Figure 4.10) and confirmed by spectral integration (Table 4.8). Therefore, the possible age-bias suspected for lipoproteins, albumin and histidine was not corroborated. On the other hand, decreases in acetate, alanine and glucose emerged as new significant changes in the age-matched subset (Table 4.8).

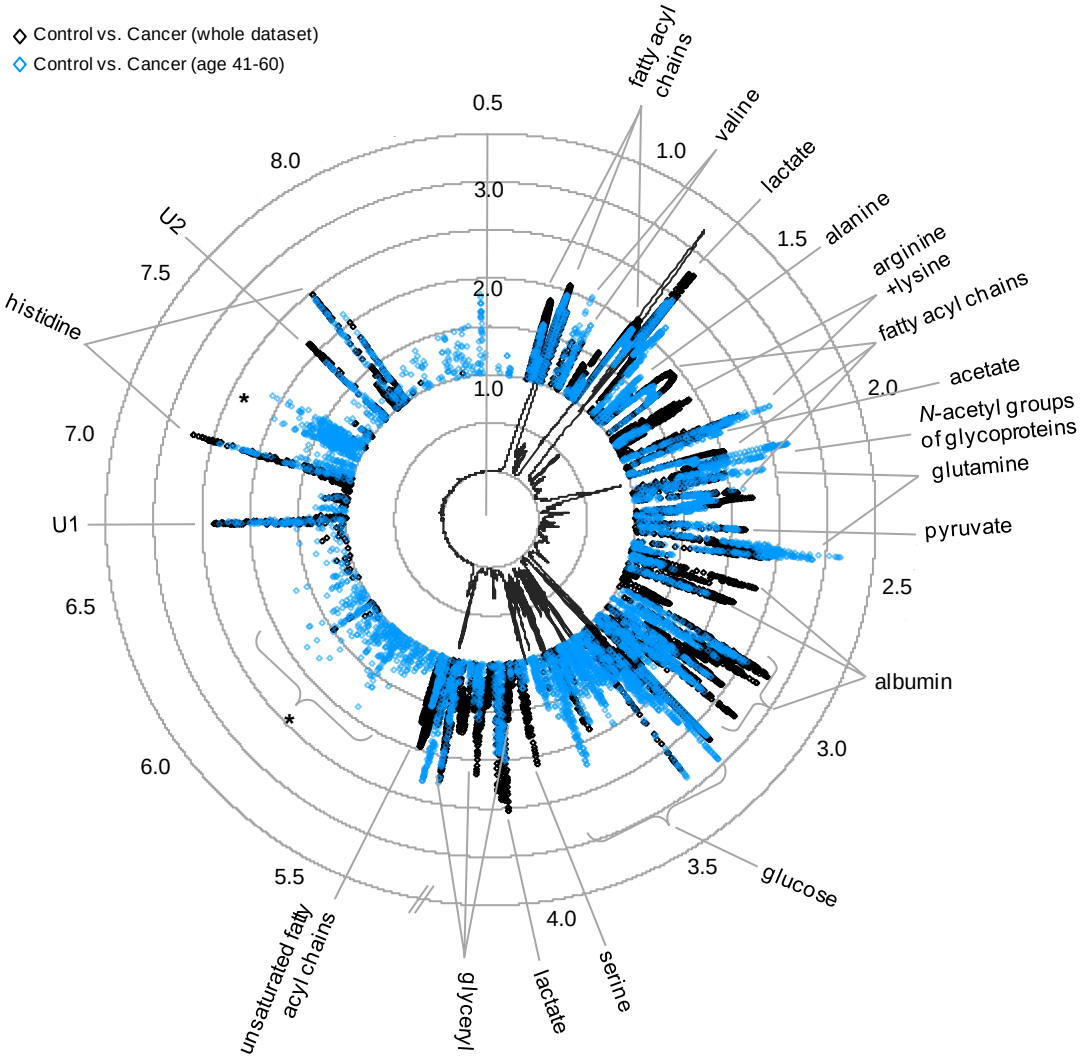


Figure 4.10 VIP-wheel representation of variables found to have VIP>1 in the plasma OPLS-DA models built with: the whole dataset (black), a subset with improved age-matching between controls and patients (blue).

4.5.3 Possible impact of smoking habits and other potential confounders

As a first approach to assess the impact of smoking habits on the plasma composition, control subjects were divided into smokers (n 29) and never-smokers (n 44) and their spectra subjected to PLS-DA. As shown in Table 4.6, a negative median Q^2 and low classification rate (58.2%) were obtained, clearly showing that the NMR plasma profiles from healthy never-smokers could not be discriminated from those of control smokers, thus hindering any further assumption about putative plasma metabolites related with smoking habits. This finding differed from the results published by Xu et al. whereby the effects of smoking and smoking cessation on human serum metabolic profile were assessed by untargeted LC-MS. The authors reported twenty-one smoking-related metabolites, including amino acids (serine, arginine, aspartate, ornithine and glutamate) and phospholipids (phosphatidylcholines, lysophosphatidylcholines and hydroxysphingomyelin), and a reversible behaviour of the serum profile after smoking cessation (Xu et al. 2013). The discrepancy between this study and ours may eventually relate to the different sensitivities of the analytical techniques used or the different composition of the groups compared (e.g., in terms of age, cigarette consumption).

With respect to the influence of smoking habits on cancer vs. control discrimination, the heterogeneous nature of the control group, comprising smokers (n 29), former smokers (n 21) and never-smokers (n 44), is expected to reduce the possibility of smoke-related bias. Still, subgroups of control subjects, divided according to their smoking habits, have been modelled by PLS-DA together with cancer patients (using also a subset of patients to avoid large imbalance in sample numbers between groups). The corresponding MCCV results are shown in Table 4.6. Compared to the whole dataset model, these new models showed similar predictive power, sensitivity and specificity, thus corroborating the low influence of smoking habits. Indeed, the corresponding loadings (not shown) were identical to those explaining the separation between control and cancer in the total model.

Concerning other potential confounders, the dietary influence was tentatively minimized by collecting all samples after overnight fasting, thus excluding diurnal variations (Park et al. 2009). Nevertheless, the effect of diet on the metabolic content of plasma could not be entirely excluded as subjects were not under any controlled/standardised dietary regime, and it is known that diet can induce significant

alterations, for instance, in the fatty acid composition of plasma (Skeaff et al. 2006). Other possible confounding factors which could account for metabolic variations, but were not evaluated in this study, are the body mass index (BMI) and weight loss or malnutrition. Lawton at al. reported high BMI to be associated with increased plasmatic creatine, kynurenine and urea (Lawton et al. 2008). Although in the present study none of these metabolites has been proposed to be cancer-related, future studies should entail the possible influence of varying BMI. In regard to weight loss, a common condition in patients with advanced stage cancer, it has been reported that alterations in glucose and alanine metabolism, detected in the plasma of lung cancer patients, may depend on the degree of weight loss (Leij-Halfwerk et al. 2000). Although knowing that most patients enrolled in our study were at stage I (where weight loss is less common), this parameter should also be assessed in future studies.

4.6 Preliminary external validation of plasma-based classification models

An important step in the validation of a biomarker compound or profile consists of testing its ability to predict class membership of an external (independent) set of samples (Bleeker et al. 2003). Despite no independent set was available, external validation was tested by using a subgroup of samples as prediction set (control n 20, cancer n 20), while the training set was composed of the remaining samples (control n 74, cancer n 86). Samples chosen to constitute the prediction set were the ones collected and analysed more recently. Class prediction of these samples was performed based on PLS-DA models built with the training set and using either the full resolution data or the twelve metabolites previously found to be more relevant in cancer vs. control discrimination. The corresponding scores scatter plots are shown in Figure 4.11a and 4.11b and the prediction results are presented in Table 4.10. When using the full resolution spectra only, 13 out of 20 controls (specificity of 65%) and 15 out of 20 cancer patients (sensitivity of 75%) were correctly classified, whilst when using the twelve relevant integrals, sensitivity and specificity increased to 85%. This result demonstrated that the amount of information contained in a full resolution spectrum may not all correlate with cancer vs. control discrimination, and that the extraction of the most relevant features can, in some instances, improve the classification accuracy.

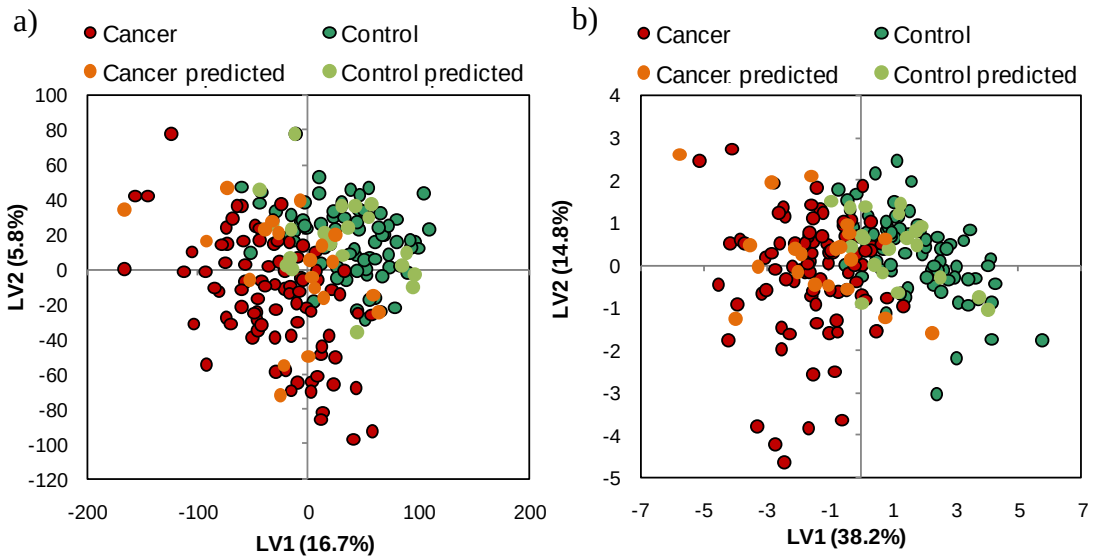


Figure 4.11 Scores scatter plot obtained by PLS-DA of a) full resolution ^{1}H NMR plasma spectra (*LV*=2, R^2X=0.23, R^2Y=0.65, Q^2=0.53), and b) 12 signal integrals (*LV*=2, R^2X=0.53, R^2Y=0.89, Q^2=0.87), using in the calibration set 74 samples from controls and 86 samples from patients. The scores corresponding to class prediction of additional 40 samples (20 controls and 20) are superimposed in lighter colour.

Table 4.10 Confusion matrix showing the prediction results for an independent test set (20 controls and 20 patients) obtained by PLS-DA modelling of a calibration set (74 controls and 86 patients), using either the full resolution plasma spectra or 12 selected signal integrals (acetoacetate, glutamine, histidine, lactate, methanol, glycoproteins, pyruvate, U1 and U2, valine, HDL and LDL+VLDL). Shaded boxes show the number of samples correctly classified.

			Full resolution spectra		12 Integrals				
			True class		True class				
			Cancer	Control	Cancer	Control			
Full resolution spectra	Predicted class	Cancer	15	7	17	3	Cancer	Predicted class	12 Integrals
		Control	5	13	3	17	Control		

4.7 Proposed biochemical interpretation of cancer-related metabolic variations in blood plasma

According to the results presented in the preceding sections, the metabolic composition of the patients' plasma differed significantly from controls in the following main aspects: increased levels of lactate, pyruvate, acetoacetate, LDL+VLDL and glycoproteins, together with decreased levels of several amino acids (glutamine, histidine, valine, serine, arginine/lysine), HDL lipoproteins, methanol and two unknowns.

The increased lactate levels in the plasma of cancer patients corroborates the glycolytic shift in energy metabolism, already observed through direct tissue analysis (subchapter 3.5), and agrees with literature reports on lung cancer (Hori et al. 2011; Jordan et al. 2010) and other cancer types, namely kidney (Gao et al. 2008), liver (Gao et al. 2009) and colorectal (Qiu et al. 2009) cancers. Moreover, the increase in pyruvate (precursor of lactate) and the trend for glucose depletion (which reached statistical significance when comparing groups with improved age matching) further agree with enhanced glycolysis.

A number of amino acids, namely glutamine, histidine, valine, serine, arginine/lysine, and alanine in SqCC, were found to be depleted in the plasma of cancer patients compared to healthy controls. Although one study reported an opposite variation for serine and alanine (Maeda et al. 2010), most literature studies are consistent in reporting decreased levels of circulating plasma free amino acids (PFFA) in several cancer types (Lai et al. 2005), including lung cancer (Miyagi et al. 2011; Shingyoji et al. 2013; Wen et al. 2013). In fact, the measurement of PFAA levels is at the basis of a diagnosis methodology being developed in Japan, the 'AminoIndex technology' (Okamoto 2012). The observed decreased in plasma amino acids may reflect their glucogenic conversion to substrates of the TCA cycle, in order to sustain the higher energetic and biosynthetic demands of tumour cells. For instance, glutamine and histidine can be converted to α-ketoglutarate, valine to succinyl-CoA, and arginine to fumarate (Figure 4.12). The decrease in plasmatic glutamine, in particular, has been observed across several cancer types, such as renal (Gao et al. 2008; Zira et al. 2010) and pancreatic (Urayama et al. 2010) cancers, and has been associated with the upregulation of glutaminolysis in cancer. Indeed, our results of tumour tissue profiling (subchapter 3.5) also corroborate this finding.

Acetoacetate, together with β-hydroxybutyrate in the case of AdCs, was found to be increased in the patients' plasma, which, to our knowledge, is a newly reported variation in lung cancer. Regarding other cancer types, acetoacetate was elevated in the serum of oral cancer patients (Tiziani et al. 2009) and decreased in renal and liver cancer subjects (Gao et al. 2008; H. C. Gao et al. 2009). Acetoacetate and β-hydroxybutyrate are ketone

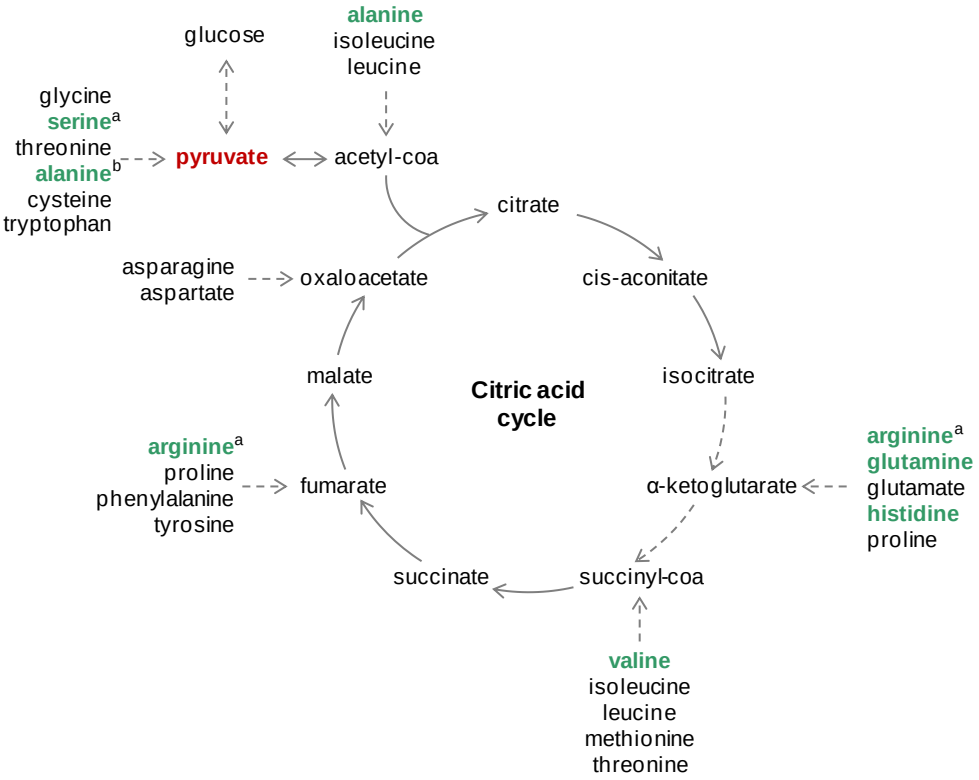


Figure 4.12 Schematic representation of the conversion of gluconeogenic amino acids to substrates of the TCA cycle. Metabolites in red were found increased in the plasma of lung cancer patients, whereas metabolites in green were found decreased in relation to controls. [a] Variations assessed qualitatively; [b] significant variation in SqCC (not in AdC).

bodies, the production of which typically occurs in response to low blood glucose availability, in order to use energy from fatty acids (Nelson and Cox 2004). The beta-oxidation of fatty acids forms acetyl-CoA, which can enter the TCA cycle for energy production, or in the case of excessive accumulation, can be used for synthesis of ketone bodies (Figure 4.13). Interestingly, recent studies have started to bring to light a relevant role for fatty acid oxidation in cancer, namely as an alternative source of ATP and reducing power (NADPH) in conditions of metabolic stress, thus providing tumour cells survival advantages (Carracedo et al. 2013).

The proportions of the main lipoprotein subclasses, namely HDL and LDL+VLDL were also found to be significantly altered in the plasma of lung cancer patients. Lipoproteins are complex structures that function as transport vehicles for water-insoluble lipids in the blood. In general, they have a spherical structure with a hydrophobic core of non-polar triglyceride and cholesterol ester molecules, surrounded by an amphiphilic surface of apoliproteins and phospholipids (Ala-Korpela 1995). Previous studies on the

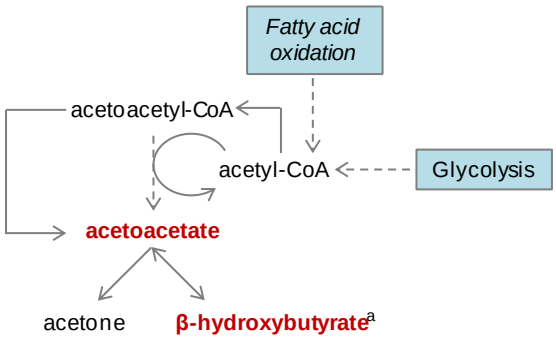


Figure 4.13 Schematic representation of the synthesis of ketone bodies (acetoacetate, acetone and β-hydroxybutyrate). Metabolites in red were increased in the plasma of lung cancer patients, blood. [a] Significant variation in AdC (not in SqCC).

serum lipoprotein profile in cancer subjects suggested significant alterations for a variety of haematological and solid tumours, including lung cancer (Fiorenza et al. 2000; Muntoni et al. 2009). In agreement with our results, Muntoni et al. reported significantly decreased HDL levels in over 500 serum samples from patients with different types of solid tumours compared to age- and gender-matched healthy controls. Additionally, the same authors reported increased triglycerides and decreased LDL levels in the patietns' serum. Concordantly, triglyceride-rich lilpoprotein (VLDL) were also found increased in the present work, whereas the variation of LDL could not be evaluated on its own due to spectral overlap. Although the biochemical origin of lipoprotein changes is not well understood, a possible explanation for decreased HDL levels is the increased uptake of circulating lipoproteins to supply the necessary cholesterol for membrane build-up of proliferating tumour cells. Moreover, it has been suggested that the general poor nutritional status of cancer patients may also account for an altered lipoprotein profile (Fiorenza et al. 1996), although not explaining it entirely (Fiorenza et al. 2000).

Another significant alteration was the increase in *N*-acetylated glycoproteins detected in the patients' plasma, similarly to the findings reported by other metabolomic studies of pancreatic (Zhang et al. 2012) and hepatocellular cancers (Nahon et al. 2012). Protein glycosylation is a common post-translational modification, known to play fundamental roles in numerous biological processes and in the pathogenesis of several diseases, including cancer (Kim and Misek 2011). Indeed, cancer cells are known to express aberrant glycosylation patterns and several serological cancer biomarkers used clinically (such as prostate-specific antigen – PSA, cancer antigen CA15.3, and carcinoembryonic antigen – CEA) are glycoproteins. This has in fact stimulated intensive

research in the field of glycoproteomics, with several studies having addressed the serum glycoproteome of lung cancer patients (Zeng et al. 2010; Tran et al. 2008; Heo et al. 2007). Thus, in the continuity of this work, it would be very interesting to further characterize the glycoprotein fraction detected by NMR through other techniques with a higher resolving power, namely using mass spectrometry proteomics.

Finally, compared to controls, lung cancer patients showed relatively lower methanol levels. Endogenous methanol has been reported to the present in the blood plasma of healthy individuals (Haffner et al. 1996; Psychogios et al. 2011), but no previous record of its variation in cancer patients was found. Being of microbial origin, it has been reported in saliva (Takeda et al. 2009) and urine of *Pneumoniae*-infected patients (Slupsky et al. 2009), but no explanation for the decrease of methanol levels in the blood plasma of cancer patients could be proposed at this point.

5 NMR METABOLOMIC STUDY OF URINE TO ASSESS METABOLIC ALTERATIONS RELATED TO LUNG CANCER

This chapter describes the results of applying NMR metabolomics to unveil urinary cancer-related metabolic features. Following a brief account on the urine composition viewed by NMR, the discrimination between lung cancer patients and healthy controls through multivariate analysis is presented, and the putative discriminant metabolites are highlighted, considering also their dependency on histological type and stage. In a subsequent section, the influence of potential confounders (gender, age, smoking habits) is evaluated. Finally, tentative biochemical interpretation for some of the main cancer-related urinary changes is presented.

5.1 Metabolic composition of human urine: spectral assignment based on 1D and 2D NMR experiments

Urine is an aqueous mixture of numerous compounds, mainly small metabolites, with variable concentration levels, giving rise to highly complex 1D ^{1}H NMR spectra with hundreds of overlapping peaks (Figure 5.1). Hence, spectral assignment of most urine metabolites was only possible with the help of 2D NMR experiments, where overlap is minimized and the identification of complete spin systems facilitated. ^{1}H-^{1}H TOCSY and ^{1}H-^{13}C HSQC experiments were used for this purpose (together with *J*-resolved) and are illustrated in Figure 5.2.

Inspection of the TOCSY spectrum allowed the spin systems of several amino acids and organic acids to be unequivocally identified, whereas the HSQC and *J*-resolved experiment were especially helpful in the assignment of several singlets which, having no scalar couplings, did not show crosspeaks in the TOCSY spectrum. For instance, a number of singlets in the 1.9-2.1 region were tentatively assigned to *N*-acetylated metabolites, although their detailed identification was not possible due to low signal intensity and lack of information on the remaning part of the molecule (STOCSY was applied using these singlets as driver peaks, but showed no relevant correlations to other resonances).

According to a literature report on the NMR analysis of urine from patients with inborn errors of metabolism, where *N*-acetylated metabolites tend to be exacerbated, these signals correspond to a range of *N*-acetylated amino acids, *N*-acetylneuraminic acid or *N*-acetylated oligosaccharides (Engelke et al. 2004). Other tentative assignments, based on previous literature reports, regarded bile acids (Trump et al. 2006), 4-deoxyerythronate and 4-deoxythreonate (Appiah-Amponsah et al. 2009; Ellis et al. 2012) and medium-chain fatty acids (Holmes et al. 1997).

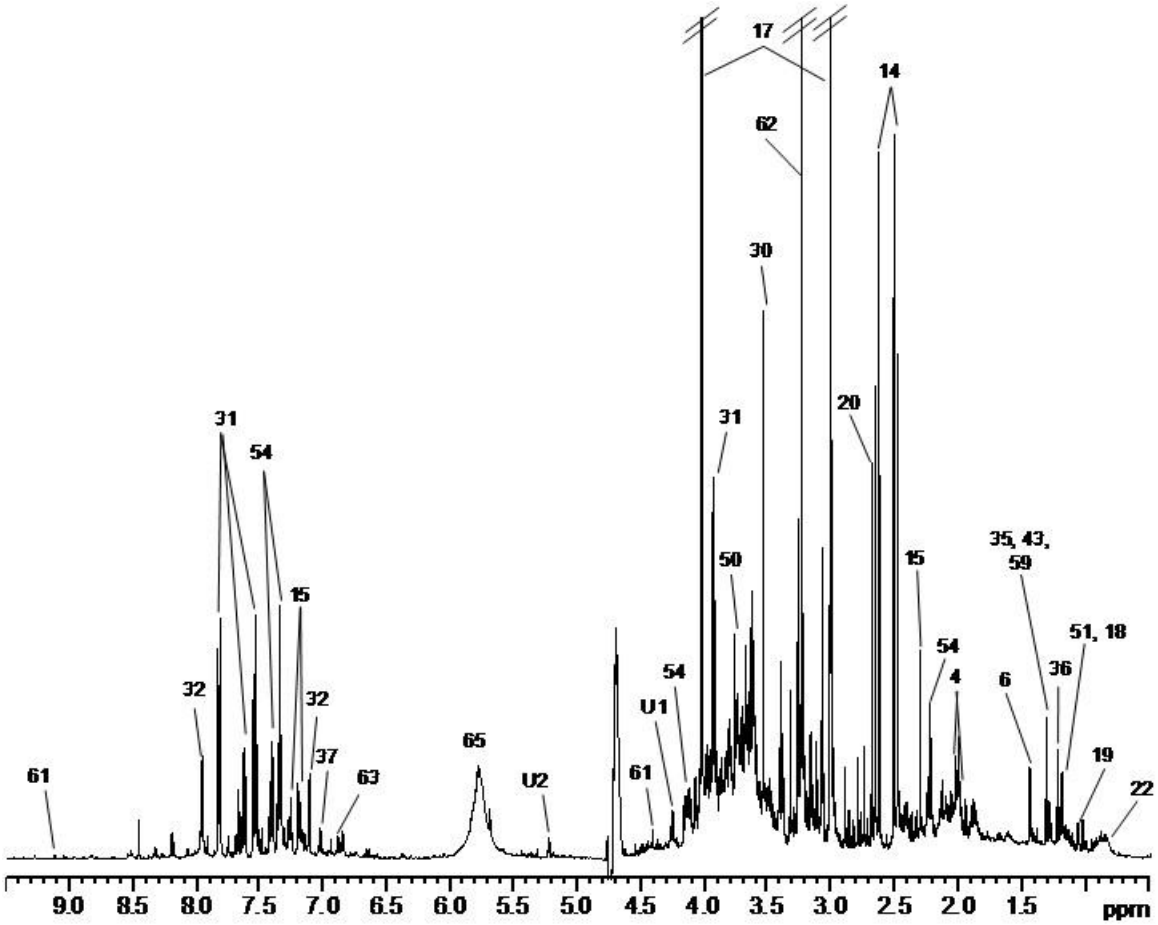


Figure 5.1 Typical 500 MHz standard 1D ^{1}H NMR spectra of urine from a lung cancer patient used for spectral assignment. Metabolites are numbered in accordance with Table 5.1. (U1: unknown 1, δ 4.30; U2: unknown 2, δ 5.35).

Moreover, spiking experiments were also performed in order to confirm the assignment of the following metabolites: 1-methylhistidine, 3-methylhistidne, pyruvate, *scyllo*-inositol, trimethylamine-*N*-oxide and betaine. Overall, based on matching chemical shifts, multiplicities and coupling constants of urine 1D and 2D spectra with those of reference compounds available in Bruker's BBIOREFCODE-2-0-0 database (Bruker

Biospin, Rheinstetten, Germany), as well as in other available on-line databases (Bouatra et al. 2013; Ulrich et al. 2008; Wishart et al. 2007) and literature reports (Holmes et al. 1997), sixty six metabolites, all of them previously reported, were identified in human urine (Table 5.1), and twenty five resonances remained unassigned.

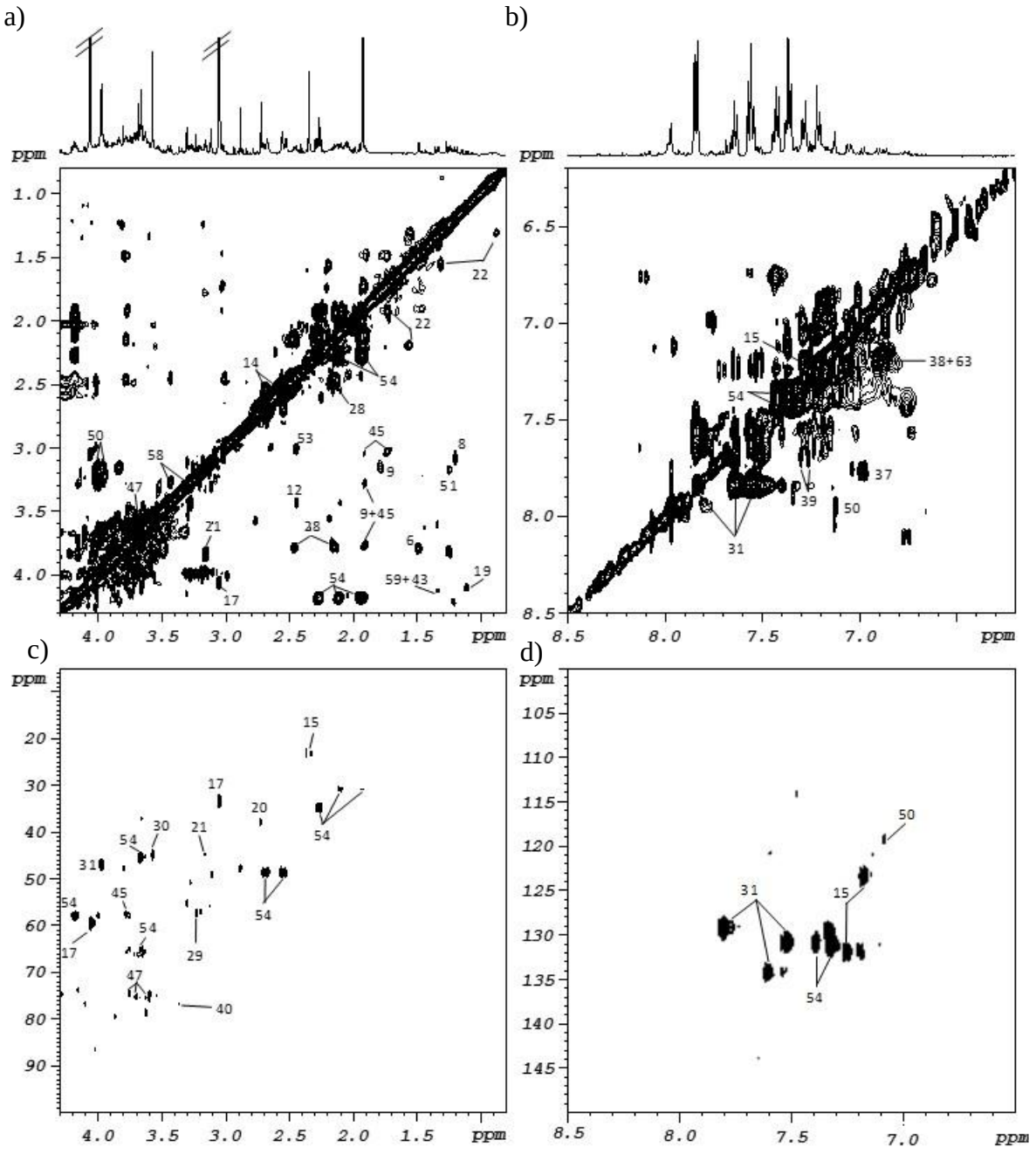


Figure 5.2 Expansions of 2D NMR spectra of urine of a lung cancer patient, used for spectral assignment: a) and b) ^{1}H-^{1}H TOCSY, c) and d) ^{1}H-^{13}C HSQC. Metabolites are numbered in accordance with Table 5.1.

Table 5.1 Assignment of resonances in the 500 MHz ^{1}H NMR spectra of human urine; (s, singlet; d, doublet; t, triplet; q, quartet; m, multiplet; dd, doublet of doublets; br, broad).

No.	Compound	δ ^{1}H in ppm (multiplicity, assignment) / δ ^{13}C in ppm
1	Acetate	1.92 (s, βCH_3)/30.71
2	Acetoacetate	2.29 (s, γCH_3)
3	Acetone	2.24 (s, CH_3)
4	*N*-Acetylated metabolites	1.99 (s, CH_3)/25.49 2.01 (s, CH_3)/25.20 2.03 (s, CH_3)/24.84 2.05 (s, CH_3) 2.07 (s, CH_3)/25.09 2.08 (s, CH_3)/24.84
5	*cis*-Aconitate	3.12 (d, CH); 5.72 (t, CH_2)
6	Alanine	1.48 (d, βCH_3)/18.92; 3.79 (q, αCH)
7	Allantoin	5.39 (CH)/66.21
8	β-Aminoisobutyrate	1.19 (d, CH_3); 2.61 (m, αCH_2); 3.07 (m, βCH_2)
9	Arginine	1.75 (γCH_2); 1.91 (βCH_2); 3.27 (t, δCH_2); 3.77 (αCH)
10	Betaine[a]	3.27 (s, CH_3)/56.16; 3.90 (s, CH_2)
11	Bile acids	0.54, 0.57 (s, C18-CH_3)
12	Carnitine	2.23 (s, CH_3)/57.19; 2.44 (αCH_2); 3.42 (γCH_2)
13	Choline	3.21 (s, $N(CH_3)_3$)/56.86; 3.52 (m, CH_2(NH)); 4.06 (m, CH_2(OH))
14	Citrate	2.54, 2.69 (dd, α,βCH_2)/48.84
15	*p*-Cresol sulphate[b]	2.35 (s, CH_3)/22.88; 7.21 (C2H, C6H)/132.89; 7.28 (C3H, C5H)
16	Creatine	3.04 (s, CH_3)/39.59; 3.94 (s, CH_2)
17	Creatinine	3.05 (s, CH_3)/33.20; 4.06 (s, CH_2)/59.44
18	4-Deoxyerythronate	1.23 (d, γCH_3)
19	4-Deoxythreonate	1.11 (d, γCH_3)
20	Dimethylamine	2.73 (s, CH_3)/37.71
21	Ethanolamine	3.15 ($CH_2(NH_2)$)/44.35; 3.83 (CH_2(OH))/60.77
22	Fatty acids	0.89/0.90/0.93 (br, CH_3); 1.30/1.44 (br, -$(CH_2)_n$-); 1.55/1.57/1.60 (br, $\underline{CH_2}$-CH_2-CO); 2.17/2.20 (br, CH_2-CO)
23	Formate	8.46 (s, CH)
24	Fumarate	6.53 (s, CH)
25	2-Furoylglycine	7.70 (C2H); 7.19 (C3H); 6.65 (C4H)
26	α-Glucose[a]	3.42 (C4H); 3.54 (C2H); 3.71 (C3H); 3.77 (C6H); 3.84 (C5H); 5.25 (d, C1H)
27	β-Glucose[a]	3.23 (C2H); 3.44 (C4H); 3.47 (C5H); 3.49 (C3H); 3.72 (C6H); 3.90 (C6'H); 4.65 (d, C1H)
28	Glutamine[a]	2.14 (m, βCH_2); 2.45 (m, γCH_2); 3.78 (αCH)

Table 5.1 (continued)

No.	Compound	δ ^{1}H in ppm (multiplicity, assignment) / δ ^{13}C in ppm
29	Glycerophosphocholine	3.23 (s, $N(H_3)_3$)/56.98
30	Glycine	3.57 (s, αCH_2)/44.59
31	Hippurate	3.97 (CH_2)/46.61; 7.56 (C2H, C6H (ring))/131.64; 7.64 (C3H, C5H (ring))/135.06; 7.83 (C4H (ring))/129.99; 8.50 (br, NH)
32	Histidine	3.17 (βCH_2); 3.28 (βэCH_2); 4.01 (αCH_2); 7.13(s, C4H, ring); 7.99 (s, C2H, ring)
33	α-Hydroxybutyrate	0.90 (t, γCH_3)/21.49; 1.65 (m, βCH); 1.72 (m, β'CH)
34	β-Hydroxybutyrate	1.19 (d, γCH_3); 2.31 (βCH); 2.42 (β'CH); 4.15 (αCH)
35	α-Hydroxyisobutyrate	1.36 (s, αCH_3)
36	β-Hydroxyisovalerate	1.27 (s, CH_3); 2.35 (βCH_2)
37	*p*-Hydroxyhippurate	3.96 (s, αCH_2)/46.72; 6.98 (C3H, C5H (ring))/119.62; 7.76 (C2H, C6H (ring))/132.45
38	*p*-Hydroxyphenyl acetate	3.45 (s, CH_2)/46.49; 6.87 (C3H, C5H (ring))/118.42; 7.17 (C2H, C6H (ring))/132.45
39	Indoxyl sulfate	7.21 (C8H)/122.67; 7.28 (C7H)/125.03, 7.37 (s, C2H)/119.01; 7.51 (C6H)/115.29; 7.71 (C9H)/120.57
40	*scyllo*-Inositol[a]	2.36 (s, CHOH)/76.63
41	Isobutyrate	1.07 (d, CH_3), 2.47 (m, CH)
42	Isoleucine	0.93 (t, δCH_3); 1.01 (d, $\beta' CH_3$)
43	Lactate	1.33 (d, βCH_3)/22.83; 4.12 (q, αCH)/68.83
44	Leucine	0.95 (d, δCH_3); 0.96 (d, $\delta' CH_3$); 3.74 (t, αCH)
45	Lysine	1.46 (m, γCH_2); 1.73 (m, δCH_2); 1.91 (m, βCH_2); 3.04 (t, εCH_2); 3.76 (t, αCH)
46	Malonate	3.11 (s, CH_2)/48.87
47	Mannitol	3.69, 3.85 (α, α'CH)/75.23; 3.76 (βCH)/78.51; 3.82 (γCH)/74.25
48	Methylguanidine	2.83 (s, CH_3)
49	1-Methylhistidine[a, b]	3.72 (s, NCH_3); 7.05 (s, C4H, ring); 7.78 (s, C2H, ring)
50	3-Methylhistidine[a, b]	3.75 (s, NCH_3); 3.28 (βCH_3); 3.98 (αCH_2); 7.15 (s, C4H, ring); 8.10 (s, C2H, ring)
51	Methylmalonate	1.24 (d, CH_3); 3.16 (q, αCH)
52	3-Methylxanthine	3.52 (s, CH_3); 8.03 (s, CH ring)
53	2-Oxoglutarate	2.45 (t, γCH_2)/33.82; 3.00 (βCH_2)
54	Phenylacetylglutamine[b]	1.93, 2.11 ($\beta\beta' CH_2$)/30.87; 2.26 (γCH_2)/34.74; 4.18 (αCH)/57.81; 3.67 (CH_2)/45.15; 7.43 (C3H, C5H, ring)/132.05; 7.36 (C3H, C5H, ring)/132.05; 7.96 (NH)
55	Pyruvate[a]	2.38 (s, CH_3)
56	Succinate	2.41 (s, CH_2)/37.06
57	Tartrate[a]	4.35 (s, CH)/76.97

Table 5.1 (continued)

No.	Compound	δ ^{1}H in ppm (multiplicity, assignment) / δ ^{13}C in ppm
58	Taurine[a]	3.27 (t, S-CH_2)/50.65; 3.43 (t, N-CH_2)/38.52
59	Threonine	1.34 (d, γCH_3); 3.59 (d, αCH); 4.26 (m, βCH)
60	Trigonellinamide	4.48 (s, N(CH_3))/51.62; 8.18 (m, C5H, ring)/148.50; 8.90 (d, C4H, ring); 8.97 (d, C6H, ring); 9.28 (s, C2H (ring)
61	Trigonelline	4.44 (s, CH_3)/51.41; 8.09 (C3H, ring)/130.73; 8.84 (C2H, C4H, ring)/148.11, 149.09; 9.12 (s, C6H, ring)/148.91
62	Trimethylamine-*N*-oxide[a]	3.28 (s, CH_3)/62.31
63	Tyrosine	6.90 (d, C3H, C5H, ring); 7.20 (d, C2H, C6H, ring)
64	Uracil	5.80 (d, C5H, ring); 7.53 (d, C6H, ring)
65	Urea	5.80 (br, NH_2)
66	Valine	0.99 (d, γCH_3); 1.04 (d, γ'CH_3); 3.62 (d, αCH)

Metabolite assignments confirmed by [a] spiking and [b] STCOSY.

5.2 Potential of urine NMR profile to discriminate between patients and control subjects

The urinary metabolic content is a reflection of the excretory and homeostatic function of urine, having potential to provide information about the metabolic adaptations to different stimuli or perturbations like a disease. Several differences were apparent between the average standard 1D ^{1}H spectra of healthy individuals (controls) and cancer subjects, for instance in the levels of hippurate, phenylacetylglutamine, *p*-cresol sulphate, trigonelline, creatine and creatinine, citrate, dimethylamine, some amino acids and *N*-acetylated metabolites (Figure 5.3). To further explore these differences and search for consistent variation patterns in the urinary profile of patients, compared to that of controls, multivariate analysis has been applied to the spectral data. Group separation, already suggested in the PCA scores scatter plot (Figure 5.4a), was clearly visible along the first latent variable of PLS-DA and OPLS-DA scores scatter plots (Figure 5.4b and Figure 5.4c, respectively). Monte Carlo cross validation (MCCV) showed this discrimination to have high predictive power (median Q^2 0.76) and classification rate (96.8%), with sensitivity and specificity values above 96% (Table 5.2). Permutation testing further confirmed the robustness of this classification, as permuted models fell along the line of no discrimination in the ROC space (Figure 5.4d) and showed a Q^2 distribution centered on

very low values and distinct from the one obtained for the original models (true classes assigned) (Figure 5.4e).

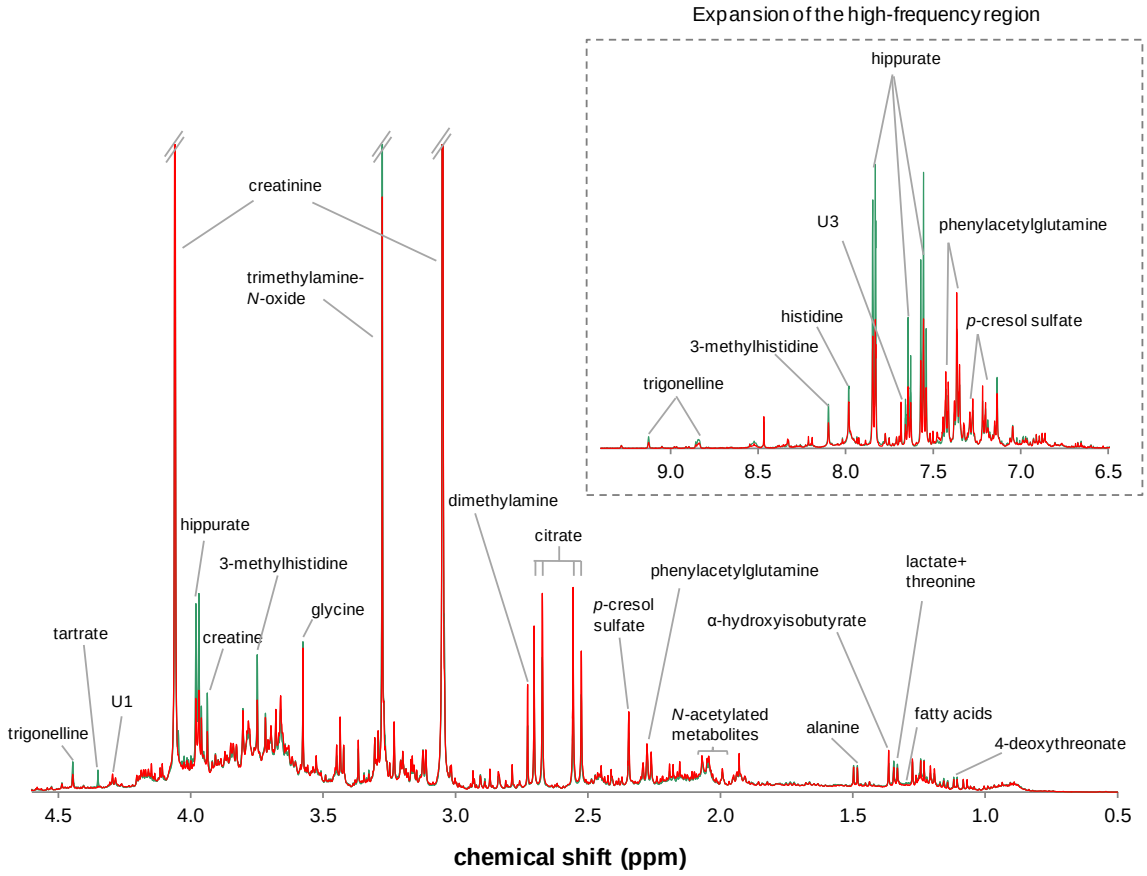


Figure 5.3 Standard 1D ^{1}H NMR average spectra of urine from controls (n 91, green) and lung cancer patients (n 109, red), after PQN normalisation. (U1: unknown 1, δ 4.30; U3: unknown 3, δ 7.68).

Table 5.2 Prediction results obtained by MCCV (500 iterations) of urine PLS-DA models assessing the discrimination between lung cancer patients and healthy controls.

PLS-DA models	Median Q^2	Sensitivity (%)	Specificity (%)	Classification rate (%)
Standard 1D	0.76	97.3	96.2	96.8
24 Integrals	0.72	97.2	94.2	95.8

The assessment of the urinary metabolites responsible for cancer vs. control discrimination was carried out by inspection of the OPLS-DA LV1 loadings plot, where each variable was coloured as a function of its VIP (variable importance in the projection)

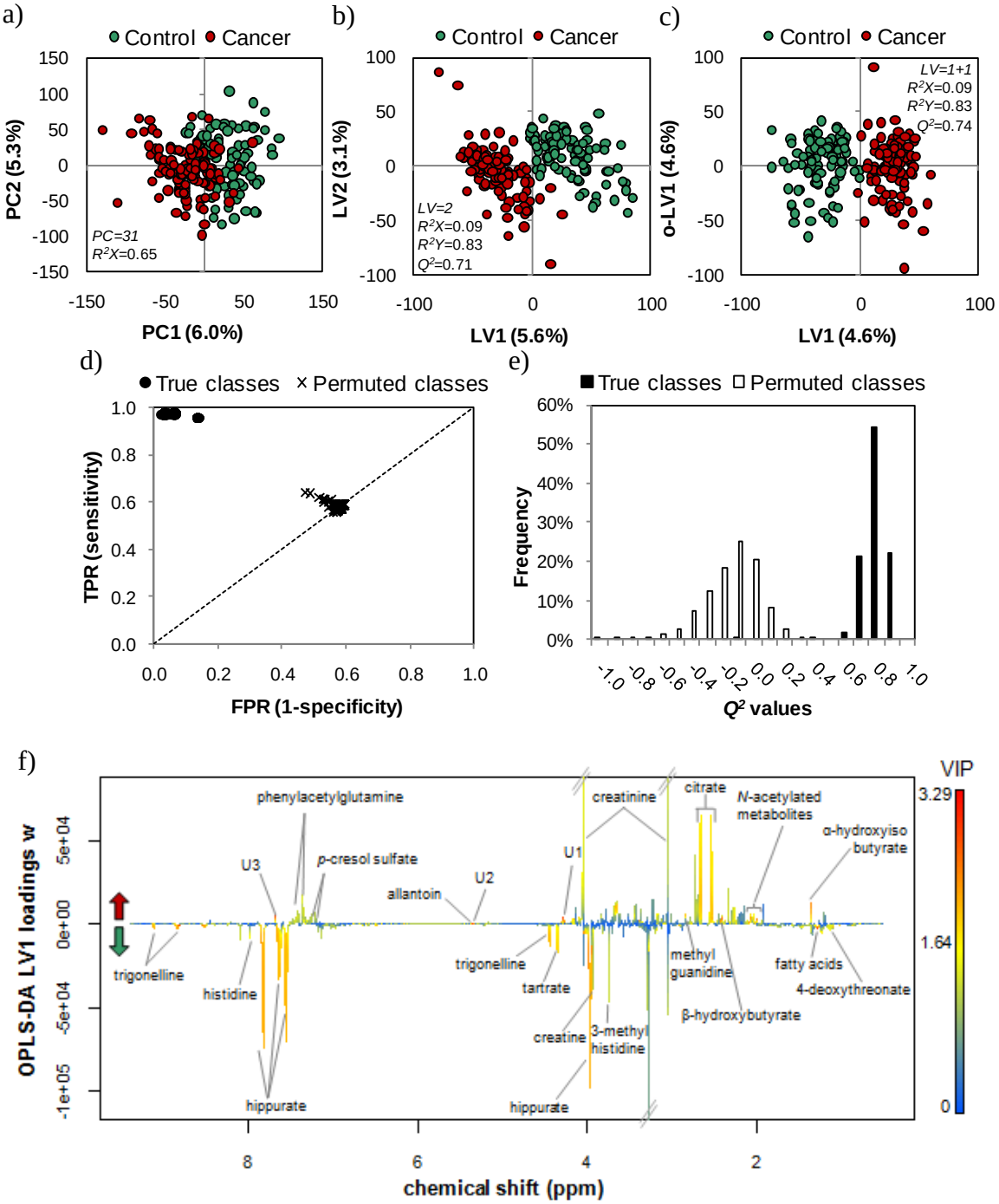


Figure 5.4 MVA applied to the standard 1D ^{1}H NMR spectra of urine from controls (n 91) and cancer patients (n 109): a) PCA, b) PLS-DA and c) OPLS-DA scores scatter plots. The parameters shown on the scores plot (*PC*: principal components, *LV*: latent variable, R^2X: variation explained by the X matrix, R^2Y: variation explained by the Y matrix, Q^2: predictive power) derive from default 7-fold cross validation. d) ROC space (TPR: true positive rate, FPR: false positive rate) and Q^2 histogram obtained by MCCV and permutation testing (500 iterations) of the PLS-DA model. f) OPLS-DA LV1 loadings weights coloured as a function of VIP. Metabolites showing VIP>1 are assigned in the plot (U1: unknown 1, δ 4.30; U2: unknown 2, δ 5.35; U3: unknown 3, δ 7.68).

(Figure 5.4f). This analysis revealed several compounds to be increased in patients' urine, namely α-hydroxyisobutyrate, several singlets in the δ 1.98-2.10 region attributed to *N*-acetylated metabolites, phenylacetylglutamine, *p*-cresol sulphate, β-hydroxybutyrate, citrate, methylguanidine, creatinine, allantoin and also three unknown resonances at δ 4.30 (multiplet), δ 5.35 (triplet) and δ 7.68 (doublet). Opposite variations (increased levels in the control group) were found for fatty acids, 4-deoxythreonate, histidine and 3-methylhistidine, creatine, tartrate, trigonelline and hippurate.

To further evaluate the significance and magnitude of the variations highlighted through loadings inspection, metabolites with VIP>1 were integrated, the median values of each group statistically compared by the Wilcoxon rank sum test and the corresponding effect size calculated. Twenty four metabolites, listed in Table 5.3, showed statistically significant differences (p<0.0013, Bonferroni-corrected) between control and cancer groups, their inter-group variations ranging from 10 to 70%. Multivariate analysis was then repeated on this set of 24 signal areas, the results being presented in Figure 5.5 and Table 5.2. Notably, this new model with a much lower number of variables behaved very well, showing MCCV parameters similar to those obtained when modelling the whole spectral range.

Table 5.3 Urine metabolites showing statistically significant differences between lung patients and healthy controls. For each metabolite, the average percentage and coefficient of variation were obtained by spectral integration of selected signals. Effect sizes and *p*-values are shown, indicating, respectively, the magnitude and statistical significance of the differences. n.s. not significant; br: broad; d: doublet; dd: double doublet; m: multiplet;.s: singlet.

Metabolite	**Cancer vs. Control**		
(δ, multiplicity)	**% variation**	***p*-value[a]**	**effect size**
Allantoin (5.39, s)	29.0±5.9	4.2×10^{-6}	0.61±0.28
Citrate (2.54, dd)	41.3±6.2	1.1×10^{-6}	0.75±0.29
Creatine (3.94, s)	-38.6±14.0	2.1×10^{-4}	-0.52±0.28
Creatinine (4.06, s)	13.8±2.7	3.1×10^{-5}	0.68±0.29
p-Cresol sulphate (2.35, s)	34.6±7.9	7.4×10^{-4}	0.51±0.28
4-Deoxythreonate (1.11, d)	-17.2±4.8	1.9×10^{-5}	-0.57±0.28
Fatty acid -CH_2- (1.30, br)	-15.3±3.2	1.8×10^{-7}	-0.79±0.29
Hippurate (7.64, m)	-50.0±12.2	2.8×10^{-11}	-0.85±0.29
Histidine (7.13, s)	-34.5±7.0	1.1×10^{-8}	-0.91±0.29
α-Hydroxyisobutyrate (1.36, s)	30.8±3.6	9.2×10^{-12}	1.02±0.30
β-Hydroxybutyrate (2.42, dd)	20.5±3.9	7.6×10^{-5}	0.64±0.28
Methylguanidine (2.83, s)	51.8±4.5	1.3×10^{-15}	1.29±0.31
N-acetylated metabolite 1 (1.99, s)	11.9±1.8	2.7×10^{-8}	0.87±0.29

Table 5.3 (continued)

Metabolite (δ, multiplicity)	Cancer vs. Control		
	% variation	*p*-value[a]	effect size
N-acetylated metabolite 2 (2.01, s)	36.4±2.3	$<2.2\times10^{-16}$	1.78±0.33
N-acetylated metabolite 3 (2.03, s)	20.1±2.3	4.0×10^{-12}	1.07±0.30
N-acetylated metabolite 4 (2.05, s)	19.9±1.7	$<2.2\times10^{-16}$	1.42±0.31
N-acetylated metabolite 5 (2.07, s)	12.5±2.4	1.1×10^{-10}	0.72±0.29
N-acetylated metabolite 6 (2.08, s)	10.7±1.8	1.7×10^{-7}	0.81±0.30
Phenylacetylglutamine (2.26, m)	20.6±4.9	2.7×10^{-4}	0.53±0.28
Tartrate (4.35, s)	-69.8±21.0	$<2.2\times10^{-16}$	-0.83±0.29
Trigonelline (9.12, s)	-49.8±11.4	2.5×10^{-10}	-0.89±0.29
Unknown (4.30, m)	24.9±2.4	$<2.2\times10^{-16}$	1.29±0.31
Unknown (5.35, m)	50.6±4.8	1.6×10^{-13}	1.16±0.30
Unknown (7.68, d)	28.3±1.9	$<2.2\times10^{-16}$	1.80±0.33

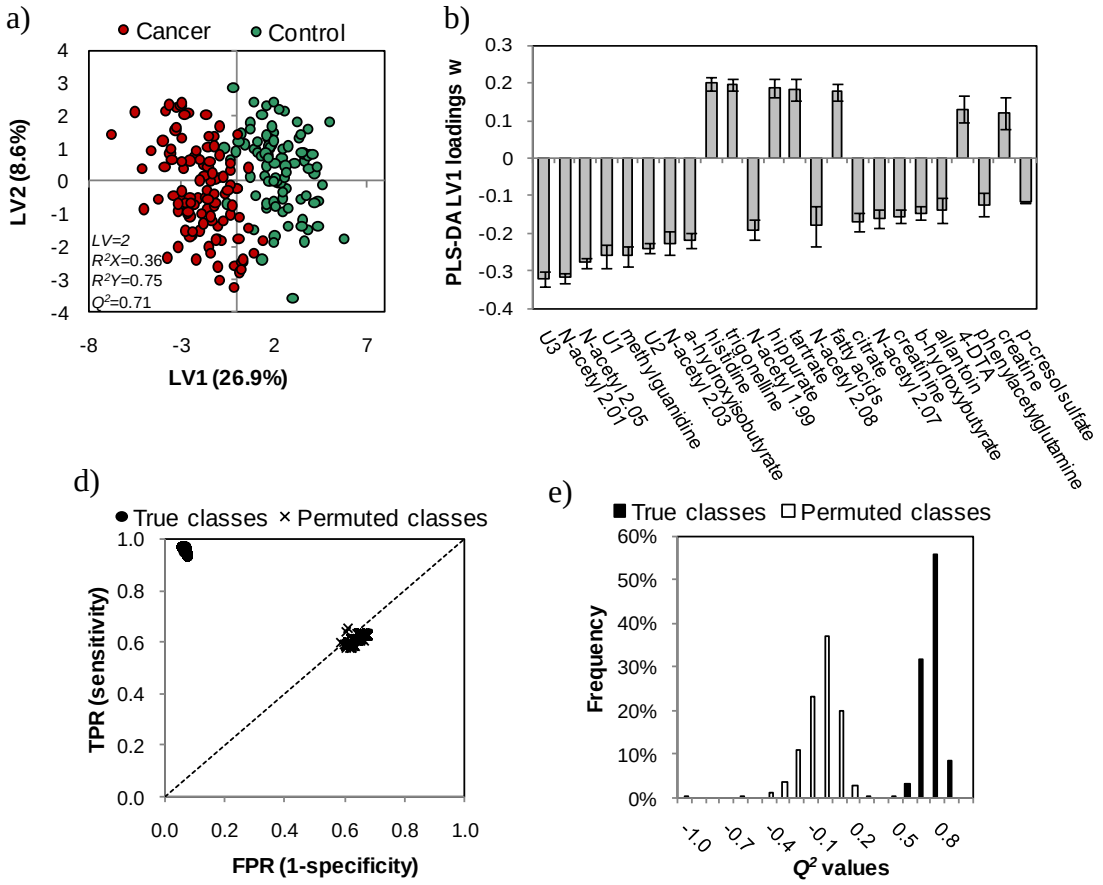


Figure 5.5 PLS-DA applied to 24 signal integrals measured in the standard 1D ^{1}H NMR spectra of urine from controls (n 91) and cancer patients (n 109): a) scores scatter plot and b) LV1 loadings. The parameters shown on the scores plot (*LV*: latent variable, R^2X: variation explained by the X matrix, R^2Y: variation explained by the Y matrix, Q^2: predictive power) derive from default 7-fold cross validation. c) ROC space (TPR: true positive rate, FPR: false positive rate) and Q^2 histogram obtained by MCCV and permutation testing (500 iterations).

5.3 Impact of tumour histological type on the urinary metabolic composition

In order to find out if the cancer-related urinary alterations were dependent on tumour histological type, the profiles of either adenocarcinoma (AdC) or squamous cell carcinoma (SqCC) (the two types for which more samples were available) were compared to those of healthy controls through multivariate analysis. The resulting PLS-DA models showed equally good MCCV quality parameters for AdC and SqCC, with high predictive power (Q^2 0.68) and classification rates (>94%) (Table 5.4). Compared to the global model, sensitivity has however slightly decreased (to around 85%), probably in relation to the unbalanced number of samples in control and cancer groups (consisting of either AdC or SqCC). Indeed, when using a smaller control group, sensitivity increased to 94% in the AdC model and to 90% in the SqCC model (Table 5.4).

Table 5.4 Prediction results obtained by MCCV (500 iterations) of urine PLS-DA models assessing the discrimination between different groups of samples (including different histological types and stages). AdC: adenocarcinoma; SqCC: squamous cell carcinoma.

PLS-DA classes [**n samples**, n M/n F, median age]	**Median Q^2**	**Sensitivity (%)**	**Specificity (%)**	**Classification rate (%)**
Control vs. Cancer [**91**, 48/43, 45] vs. [**109**, 77/32, 63]	0.76	97.3	96.2	96.8
Control vs. AdC [**91**, 48/43, 45] vs. [**42**, 23/19, 65]	0.68	85.6	98.0	94.1
Control vs. SqCC [**91**, 48/43, 45] vs. [**27**, 25/2, 63]	0.68	84.7	98.3	95.2
Control vs. AdC[a] [**42**, 26/16, 47] vs. [**42**, 23/19, 65]	0.71	94.2	96.0	95.1
Control vs. SqCC[a] [**42**, 26/16, 47] vs. [**27**, 25/2, 63]	0.67	90.1	96.8	94.2
AdC vs. SqCC [**42**, 23/19, 65] vs. [**27**, 25/2, 63]	-0.24	36.2	71.8	57.9
AdC vs. SqCC[b] [**42**, 23/19, 65] vs. [**27**, 25/2, 63]	0.46	70.0	87.5	80.6
Control vs. Cancer stage I [**91**, 48/43, 45] vs. [**63**, 45/18, 63]	0.68	94.0	97.0	95.8
Stage I vs. Stage II+III [**63**, 45/18, 63] vs. [**34**, 25/8, 62]	-0.21	27.4	70.7	55.5
Stage I vs. Stage II+III[b] [**63**, 45/18, 63] vs. [**34**, 25/8, 62]	0.39	62.1	89.4	79.8

[a] Using balanced sample numbers in the groups compared. [b]After variable selection.

Regarding the metabolic features explaining the discrimination between each histological type and the control group, they were largely similar to those highlighted when considering the whole dataset, as assessed by the loadings' analysis (not shown). However spectral integration allowed some differences to be identified, as shown in Table 5.5. In particular, neither creatine nor creatinine showed significant differences when assessed in each histological type in relation to controls. Moreover, the difference in citrate, *p*-cresol sulphate, 4-deoxythreonate, β-hydroxybutyrate, phenylacetylglutamine and trigonelline was only significant when comparing AdC patients and controls, whereas the increase in allantoin was only significant for SqCC. In regard to other metabolic features, no apparent large differences were noted in their % variation.

The ability to discriminate between AdC and SqCC based on the urinary metabolic profile was then further investigated through multivariate modelling of cancer samples, leaving out the control group. When considering the whole spectral rage (over 25000 variables), class separation was apparent in the PLS-DA scores scatter plot (Figure 5.6a), but MCCV did not validate this result, as it can be seen by the superimposition of true and permuted models in the ROC space and Q^2 histogram, shown in Figure 5.6b, c. However, after applying the variable selection method described in subchapter 2.3.5 (which reduced the number of variables to about 1600), the MCCV quality parameters were significantly improved (Table 5.4 and Figure 5.6d-f). After careful inspection of the variables selected, twenty six metabolites were integrated and their corresponding *p*-values and effect sizes calculated. However, out of these, none showed statistically significance differences between AdC and SqCC urine samples. This discrepancy between multivariate and univariate analysis results suggests that the later may be inadequate to capture the subtle differences that may be hiding in the complex urinary profiles, thus requiring the more holistic multivariate approach.

Table 5.5 Urine metabolites showing statistically significant differences between lung patients and healthy controls, considering all samples in the patients group (2nd column), only adenocarcinomas (3rd comlumn) or only squamous cell carcinomas (4th column). For each metabolite, the average percentage and coefficient of variation were obtained by spectral integration of selected signals. Effect sizes and *p*-values are shown, indicating, respectively, the magnitude and statistical significance of the differences. n.s. not significant; AdC: adenocarcinoma; SqCC: squamous cell carcinoma br: broad; d: doublet; dd: double doublet; m: multiplet;.s: singlet.

Metabolite (δ, multiplicity)	**Cancer vs. Control**			**Cancer vs. Control (AdC)**			**Cancer vs. Control (SqCC)**		
	% variation	***p*-value[a]**	**effect size**	**% variation**	***p*-value[a]**	**effect size**	**% variation**	***p*-value[a]**	**effect size**
Allantoin (5.39, s)	29.0±5.9	4.2×10^{-6}	0.61±0.28		n.s.		43.2±9.3	1.3×10^{-5}	0.89±0.44
Citrate (2.54, dd)	41.3±6.2	1.1×10^{-6}	0.75±0.29	62.1±10.2	1.5×10^{-7}	1.16±0.39		n.s.	
Creatine (3.94, s)	-38.6±14.0	2.1×10^{-4}	-0.52±0.28		n.s.			n.s.	
Creatinine (4.06, s)	13.8±2.7	3.1×10^{-5}	0.68±0.29		n.s.			n.s.	
p-Cresol sulphate(2.35, s)	34.6±7.9	7.4×10^{-4}	0.51±0.28	48.0±11.9	2.4×10^{-4}	0.73±0.38		n.s.	
4-Deoxythreonate (1.11, d)	-17.2±4.8	1.9×10^{-5}	-0.57±0.28	-25.5±4.9	2.4×10^{-6}	-0.89±0.38		n.s.	
Fatty acid -CH_2- (1.30, br)	-15.3±3.2	1.8×10^{-7}	-0.79±0.29	-14.8±3.7	1.6×10^{-4}	-0.67±0.37	-17.3±4.1	7.9×10^{-5}	-0.76±0.44
Hippurate (7.64,)	-50.0±12.2	2.8×10^{-11}	-0.85±0.29	-46.6±11.9	8.1×10^{-7}	-0.67±0.37	-46.7±11.7	6.1×10^{-5}	-0.65±0.44
Histidine (7.13, s)	-34.5±7.0	1.1×10^{-8}	-0.91±0.29	-33.7±7.0	1.8×10^{-5}	-0.79±0.38	-39.4±7.8	1.5×10^{-5}	-0.88±0.44
α-Hydroxyisobutyrate (1.36, s)	30.8±3.6	9.2×10^{-12}	1.02±0.30	25.9±6.1	1.9×10^{-5}	0.88±0.38	28.0±4.8	1.3×10^{-6}	1.20±0.45
β-Hydroxybutyrate (2.42, dd)	20.5±3.9	7.6×10^{-5}	0.64±0.28	25.6±5.8	6.9×10^{-5}	0.89±0.38		n.s.	
Methylguanidine (2.83, s)	51.8±4.5	1.3×10^{-15}	1.29±0.31	57.6±6.4	8.1×10^{-11}	1.43±0.40	55.6±8.8	9.5×10^{-8}	1.34±0.46
N-acetylated metabolite 1 (1.99, s)	11.9±1.8	2.7×10^{-8}	0.87±0.29	12.4±2.2	1.8×10^{-6}	1.06±0.39	14.5±4.1	8.3×10^{-4}	1.03±0.45
N-acetylated metabolite 2 (2.01, s)	36.4±2.3	$<2.2\times10^{-16}$	1.78±0.33	34.8±3.2	6.4×10^{-7}	2.22±0.45	42.5±6.9	4.2×10^{-12}	1.99±0.50
N-acetylated metabolite 3 (2.03, s)	20.1±2.3	4.0×10^{-12}	1.07±0.30	23.1±3.6	9.6×10^{-9}	1.26±0.36	24.8±4.6	4.2×10^{-7}	1.39±0.46
N-acetylated metabolite 4 (2.05, s)	19.9±1.7	$<2.2\times10^{-16}$	1.42±0.31	21.2±2.5	2.4×10^{-12}	1.65±0.42	25.7±2.8	1.0×10^{-11}	2.16±0.51
N-acetylated metabolite 5 (2.07, s)	12.5±2.4	1.1×10^{-10}	0.72±0.29	12.8±2.9	1.5×10^{-7}	0.71±0.38	17.8±3.6	2.2×10^{-7}	0.94±0.45
N-acetylated metabolite 6 (2.08, s)	10.7±1.8	1.7×10^{-7}	0.81±0.30	10.5±2.6	8.1×10^{-5}	0.77±0.38	14.0±2.6	9.0×10^{-6}	1.09±0.45
Phenylacetylglutamine (2.26, m)	20.6±4.9	2.7×10^{-4}	0.53±0.28	24.9±6.3	1.7×10^{-4}	0.71±0.38		n.s.	
Tartrate (4.35, s)	-69.8±21.0	$<2.2\times10^{-16}$	-0.83±0.29	-73.2±16.8	1.7×10^{-12}	-0.72±0.38	-64.0±16.7	2.3×10^{-6}	-0.58±0.44
Trigonelline (9.12, s)	-49.8±11.4	2.5×10^{-10}	-0.89±0.29	-56.7±11.3	5.2×10^{-8}	-0.92±0.38		n.s.	
U1 (4.30, m)	24.9±2.4	$<2.2\times10^{-16}$	1.29±0.31	25.0±3.7	6.9×10^{-10}	1.30±0.40	29.2±4.6	2.1×10^{-8}	1.56±0.47
U2 (5.35, m)	50.6±4.8	1.6×10^{-13}	1.16±0.30	48.0±7.2	2.8×10^{-8}	1.15±0.39	62.5±10.4	2.8×10^{-7}	1.40±0.47
U3 (7.68, d)	28.3±1.9	$<2.2\times10^{-16}$	1.80±0.33	28.4±2.6	6.8×10^{-14}	1.84±0.43	29.1±3.2	2.4×10^{-10}	1.87±0.49

[a]Wilcoxon rank sum test $p<1.4\times10^{-3}$ (Bonferroni-corrected).

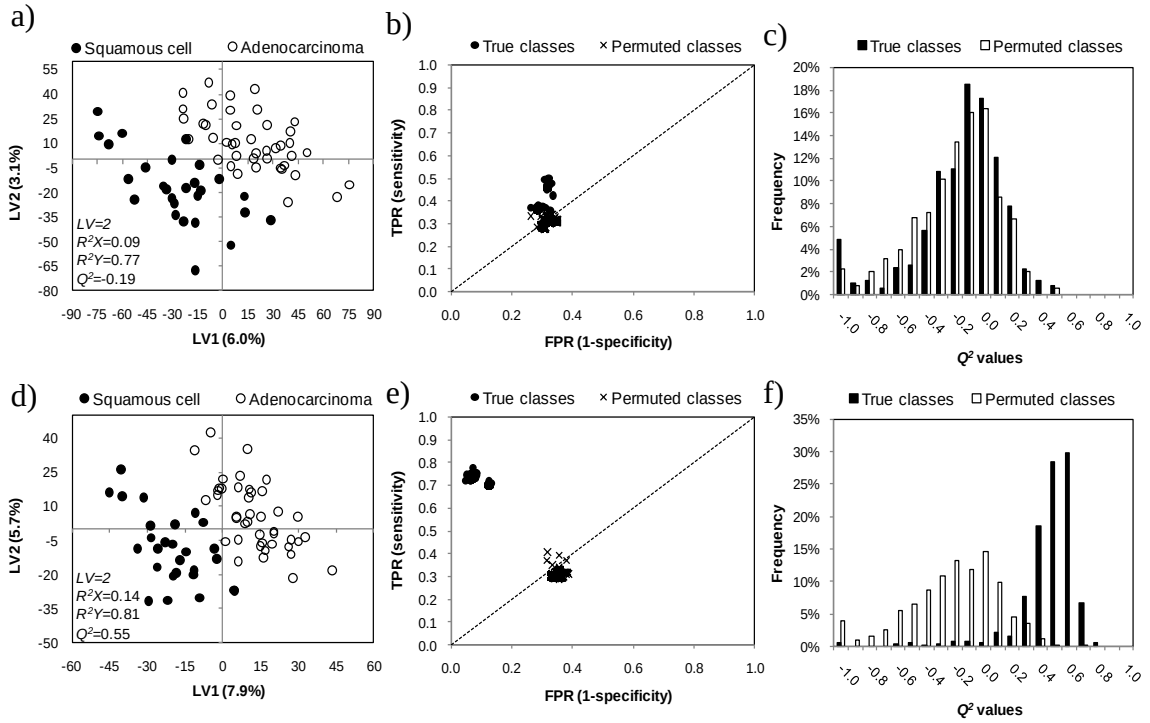


Figure 5.6 PLS-DA applied to the urine ^{1}H NMR spectra of adenocarcinoma and squamous cell carcinoma patients, either using the full resolution data (upper row) or after variable selection (lower row): a) and d) PLS-DA scores scatter plots, b) and e) ROC spaces (TPR: true positive rate, FPR: false positive rate) and c) and f) Q^2 histograms obtained by MCCV and permutation testing (500 iterations) of PLS-DA models. The parameters shown on the scores plots (*LV*: latent variable, R^2X: variation explained by the X matrix, R^2Y: variation explained by the Y matrix, Q^2: predictive power) derive from default 7-fold cross validation.

5.4 Impact of tumour stage on the urinary metabolic composition

Similarly to the analysis conducted for plasma, the possible influence of disease stage on the urinary profiles was assessed, in a first instance, by modelling each stage in comparison with the control group and, secondly, by testing pairwise comparisons between stages (leaving out the controls). Stage I patients (n 63) could be discriminated from healthy controls (n 91) with a robustness comparable to that attained when modelling the whole dataset (Table 5.4 and Figure 5.7a, b). Moreover, the analysis of the corresponding loadings showed that the variables explaining stage I vs. control discrimination were the same as those previously highlighted as important in the discrimination between all patients (regardless of disease stage) and controls (Figure 5.7c). In other words, these results show that putative cancer-related metabolic alterations are detectable in the urine of patients from early stages of cancer progress. The same approach was followed for stage II

(n 24) and stage III (n 10) samples and equally coincident signatures were obtained, although in some cases, PLS-DA models could not be properly validated, probably due to the unbalanced number of samples in the groups compared.

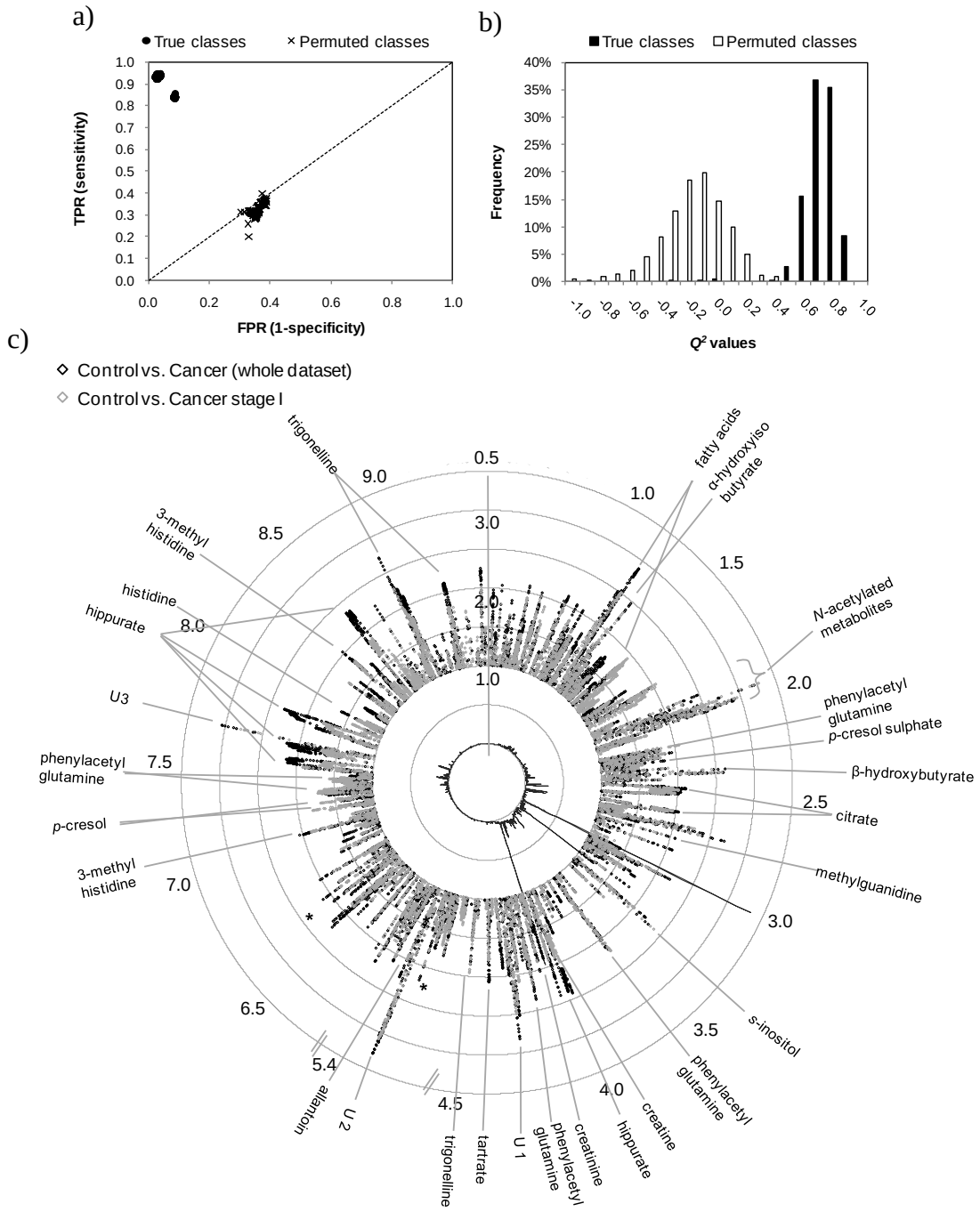


Figure 5.7 a) ROC space (TPR: true positive rate, FPR: false positive rate) and b) Q^2 histogram obtained by MCCV and permutation testing (500 iterations) of the PLS-DA model built with ^{1}H NMR spectra of urine to assess control vs. cancer stage I discrimination. c) VIP-wheel representation of variables found to have VIP>1 in the OPLS-DA models built with: the whole dataset (black), the subset with controls and stage I patients (grey).

In regard to the pairwise comparisons between cancer samples of different stages, the only model with reasonable MCCV parameters was obtained for stage I vs. stage II+III, after variable selection (Table 5.4). Still, the predictive power and classification rate were only of 0.39 and 79.8%, respectively, and further analysis of selected variables by spectral integration did not confirm any statistically significant differences. Thus, there seems to be a weak relationship between lung cancer stage and the urinary metabolic profile viewed by NMR.

5.5 Influence of potential confounders in urine-based cancer vs. control discrimination

As already mentioned in the previous chapter, the control and cancer groups available for this study were not perfectly matched in terms of age, gender and smoking habits, which, among other factors, are known to possibly influence the biofluids composition. Hence, the urinary metabolic variability associated with these factors has been investigated, as described in the following sections.

5.5.1 Gender-related metabolic features in urine

To assess the influence of gender on the urinary metabolic profile, MVA was firstly applied to spectral data from control subjects, divided into males (n 48, average age 42) and females (n 43, average age 43). PLS-DA afforded a good discrimination between genders (0.58 median Q^2 and 88% classification rate, as determined by MCCV, Table 5.6), indicating that gender has a strong impact on the metabolic profile of human urine.

Indeed, males and females were clearly separated along LV1 in the scores scatter plot (Figure 5.8a) the metabolites underlying this separation being identified in the corresponding loadings (Figure 5.8b). Females' urine was characterized by higher levels of phenylacetylglutamine, *p*-cresol sulphate, creatine, citrate, hippurate and of two unknown resonances at δ 2.41 and δ 3.80, while males' urine presented higher amounts of methylmalonate, creatinine, taurine, *s*-inositol, tartrate and of a doublet at δ 1.15.

Table 5.6 Prediction results obtained by MCCV (500 iterations) of urine PLS-DA models assessing the discrimination between different groups of samples (including groups varying in gender, age and smoking habits).

PLS-DA classes [**n samples**; n M/n F; median age]	**Median Q^2**	**Sensitivity (%)**	**Specificity (%)**	**Classification rate (%)**
Control vs. Cancer [**91**; 48/43; 45] vs. [**109**; 77/32; 63]	0.76	97.3	96.2	96.8
Male vs. Female, controls only [**48**; 48/0; 44] vs. [**43**; 0/43; 46]	0.58	88.2	87.8	88.0
Control vs. Cancer, males only [**48**; 48/0; 44] vs. [**77**; 77/0; 64]	0.79	97.7	96.8	97.3
Control vs. Cancer, females only [**43**; 0/43; 46] vs. [**32**; 0/32; 58]	0.71	91.6	96.7	94.5
Control vs. Cancer, age 41-60 [**42**; 20/22; 50] vs. [**38**; 22/16; 56]	0.70	92.6	88.2	90.3
Smoker vs. Non-smoker, controls only [**28**; 17/11; 39] vs. [**43**; 19/24; 45]	0.38	62.6	81.6	74.1
Control smoker vs. Cancer [a] [**28**; 17/11; 39] vs. [**45**; 25/20; 57]	0.70	96.3	92.2	94.7
Control non-smoker vs. Cancer [a] [**43**; 19/24; 45] vs. [**45**; 25/20; 57]	0.65	95.0	87.6	91.2

[a] Using balanced sample numbers in the groups compared.

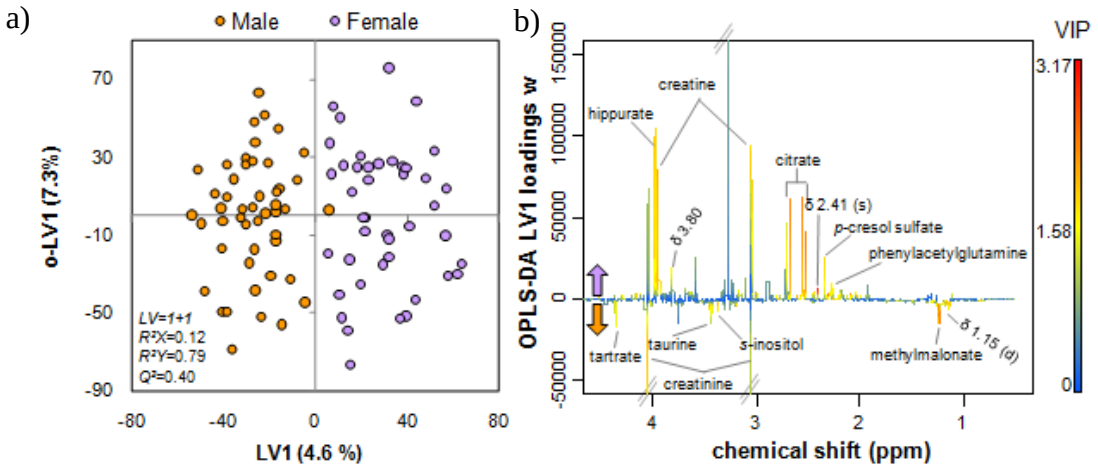


Figure 5.8 OPLS-DA applied to the ^{1}H NMR spectra of urine from control males (n 48) and females (n 43): a) scores scatter plot and b) LV1 loadings weights, coloured as a function of VIP. The parameters shown on the scores plot (*LV*: latent variable, R^2X: variation explained by the X matrix, R^2Y: variation explained by the Y matrix, Q^2: predictive power) derive from default 7-fold cross validation.

These variations were not surprising as significant gender-related differences have been previously reported in urine (Kochhar et al. 2006; Slupsky et al. 2007; Psihogios et al. 2008; Zhang et al. 2011; Swann et al. 2013). In agreement with other studies (Kochhar et al. 2006; Ramadan et al. 2006; Slupsky et al. 2007; Rasmussen et al. 2011; S. Zhang et al. 2011; Bouatra et al. 2013; Swann et al. 2013), creatinine excretion was higher in men, possibly in relation to their higher muscle mass, when compared to women. Our results

were also consistent with the literature regarding elevated levels of taurine and methylmalonate in male urine (Kochhar et al. 2006; Ramadan et al. 2006; Bouatra et al. 2013; Swann et al. 2013) and of creatine, citrate, hippurate and *p*-cresol sulphate in female urine (Slupsky et al. 2007; Swann et al. 2013). In particular, the endogenous synthesis of creatine has been proposed to be enhanced in women (Slupsky et al. 2007), while citrate levels have been related to estrogen regulation, through the control exerted by this hormone on mitochondrial and cytosolic pH and, hence, on citrate transport and excretion (Kochhar et al. 2006). The other metabolic variations found to be important in male vs. female discrimination, namely increased *s*-inositol and tartrate in males and phenylacetylglutamine in females, have not, to the best of our knowledge, been previously reported.

Amongst the above mentioned gender-related metabolic features, those which could potentially be influencing the cancer vs. control discrimination (due to gender mismatching) were creatinine (simultaneously elevated in male and cancer groups), creatine and hippurate (simultaneously reduced in male and cancer groups), as summarised in Table 5.7.

Table 5.7 Gender-related metabolic variations in urine and their comparison with cancer-related variations.

Gender-related metabolites	Variation male vs. female	Variation cancer vs. control	Observations
Citrate	↓	↑	Opposite variation
Creatine	↓	↓	Possible bias[a]
Creatinine	↑	↑	Possible bias[a]
p-Cresol sulphate	↓	↑	Opposite variation
Hippurate	↓	↓	Possible bias[a]
s-Inositol	↓	-	
Methylmalonate	↓	-	
Phenylacetylglutamine	↓	↑	Opposite variation
Tartrate	↑	↓	Opposite variation
Taurine	↑	-	
U1 (δ 1.15, d)	↑	-	
U2 (δ 1.25, d)	↑	-	
U1 (δ 2.41, s)	↓	-	

[a] Metabolites increased/decreased in both male and cancer groups may potentially bias cancer vs. control discrimination as the cancer group comprises ~70% male subjects.

In order to evaluate the true weight of these metabolites as potential cancer markers, multivariate modelling of control and cancer groups was performed separately for each gender. The resulting MCCV-validated PLS-DA models afforded, in both cases, high classification rates and predictive powers, thus, confirming the discrimination between

patients and controls to be equally good for males and females (Table 5.6). Regarding the cancer-related metabolic features highlighted in each OPLS-DA model, these were very similar to those seen when considering the whole dataset, as shown by the high degree of overlap between variables with VIP>1 in the VIP-wheel representation of the three models: cancer vs. control (n 200, whole dataset), cancer vs. control (n 125, males only) and cancer vs. control (n 75, females only) (Figure 5.9). However, some differences are worth of mention. When assessed within each gender, the difference in creatine and creatinine levels between patients and controls was no longer significant (Table 5.8), thus confirming their gender-related bias. Hippurate, on the other hand, was confirmed to be important in cancer vs. control discrimination regardless of gender. Still of notice is the fact that some differences are more relevant in one gender than the other: 4-deoxythreonate is significantly different only between female controls and patients, whereas the difference in the levels of allantoin, citrate, *p*-cresol sulphate, fatty acids, histidine, β-hydroxybutyrate, one *N*-acetylated metabolite and phenylacetylglutamine is more pronounced between males (Table 5.8).

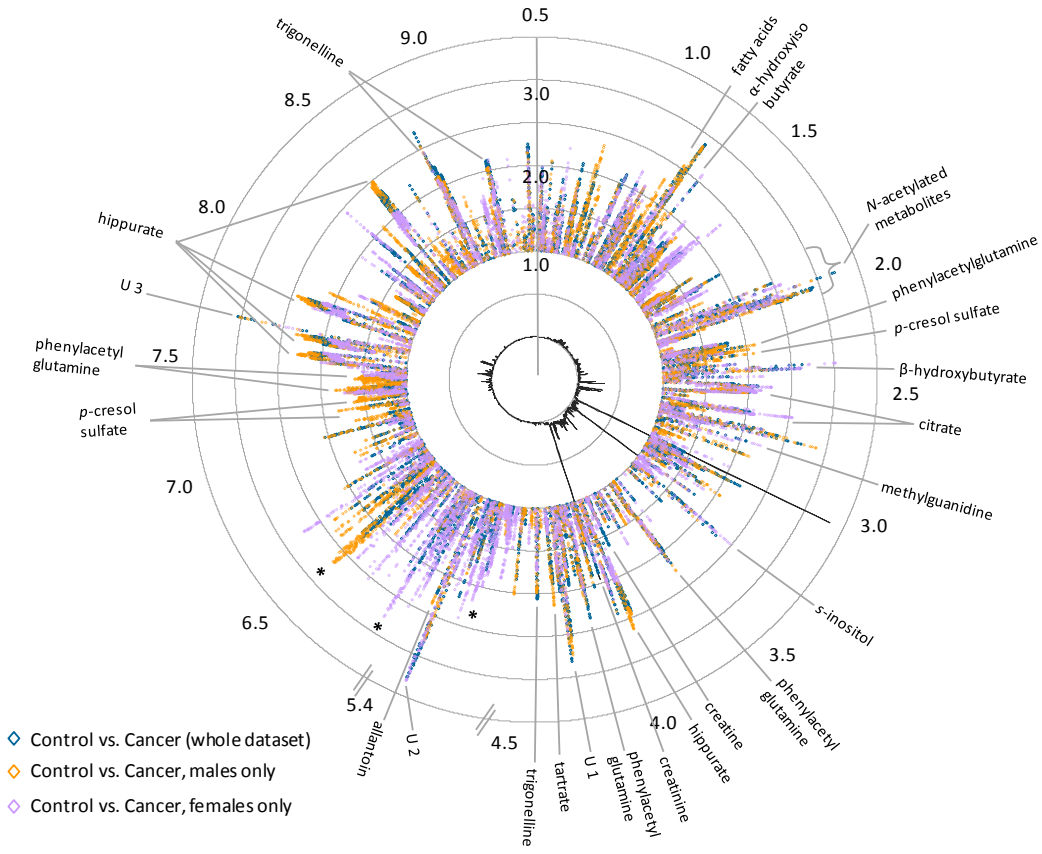


Figure 5.9 VIP-wheel representation of variables found to have VIP>1 in the urine OPLS-DA models built with: the whole dataset (blue), male subjects only (orange) and female subjects only (purple).

Table 5.8 Urine metabolites showing statistically significant differences between lung cancer patients and healthy controls, considering all samples in the patients group (2nd column), only males (3rd column), only females (4th column) or an age-matched subset (5th). For each metabolite, the average percentage and coefficient of variation were obtained by spectral integration of selected signals. Effect sizes and *p*-values are shown, indicating, respectively, the magnitude and statistical significance of the differences between the two groups. n.s. not significant; ↓ qualitative decreased (signal not integrated due to spectral overlap); s: singlet; d: doublet; dd: double doublet; m: multiplet; br: broad.

Metabolite (δ, multiplicity)	Cancer vs. Control (ALL SAMPLES)			Cancer vs. Control (MALES ONLY)			Cancer vs. Control (FEMALES ONLY)			Cancer vs. Control (AGE-MATCHED)		
	% variation	*p*-value[a]	effect size	% variation	*p*-value[a]	effect size	% variation	*p*-value[a]	effect size	% variation	*p*-value[a]	effect size
Allantoin (5.39, s)	29.0±5.9	4.2×10^{-6}	0.61±0.28	24.9±7.7	5.7×10^{-4}	0.54±0.37		n.s.		45.6±9.0	4.1×10^{-5}	0.94±0.46
Citrate (2.54, dd) [c]	41.3±6.2	1.1×10^{-6}	0.75±0.29	57.6±7.6	7.8×10^{-6}	0.90±0.38		n.s.			n.s.	
Creatine (3.94, s) [b]	-38.6±14.0	2.1×10^{-4}	-0.52±0.28		n.s.			n.s.			n.s.	
Creatinine (4.06, s) [b]	13.8±2.7	3.1×10^{-5}	0.68±0.29		n.s.			n.s.		19.2±4.1	1.7×10^{-4}	0.96±0.29
p-Cresol sulphate(2.35, s)	34.6±7.9	7.4×10^{-4}	0.51±0.28	55.8±9.0	7.1×10^{-5}	0.77±0.37		n.s.			n.s.	
4-Deoxythreonate (1.11, d)	-17.2±4.8	1.9×10^{-5}	-0.57±0.28		n.s.		-27.0±7.2	6.8×10^{-4}	-0.92±0.48		n.s.	
Fatty acid -CH_2- (1.30, br)	-15.3±3.2	1.8×10^{-7}	-0.79±0.29	-20.8±4.7	1.6×10^{-7}	-1.07±0.38		n.s.			n.s.	
Hippurate (7.64, m)	-50.0±12.2	2.8×10^{-11}	-0.85±0.29	-38.2±9.1	1.5×10^{-7}	-1.03±0.38	-54.4±18.1	1.6×10^{-4}	-0.82±0.48		n.s.	
Histidine (7.13, s) [c]	-34.5±7.0	1.1×10^{-8}	-0.91±0.29	-39.9±9.7	6.5×10^{-8}	-1.17±0.39		n.s.			n.s.	
α-Hydroxyisobutyrate (1.36, s)	30.8±3.6	9.2×10^{-12}	1.02±0.30	26.0±4.6	6.2×10^{-6}	0.83±0.38	33.6±5.6	1.2×10^{-6}	1.30±0.50	40.7±5.6	1.8×10^{-8}	1.39±0.49
β-Hydroxybutyrate (2.42, dd)	20.5±3.9	7.6×10^{-5}	0.64±0.28	24.5±4.1	3.0×10^{-5}	0.83±0.37		n.s.		28.4±6.2	1.0×10^{-4}	0.93±0.46
Methylguanidine (2.83, s)	51.8±4.5	1.3×10^{-15}	1.29±0.31	40.7±5.0	6.4×10^{-9}	1.15±0.39	71.0±8.4	1.8×10^{-8}	1.50±0.51	51.9±7.4	1.3×10^{-7}	1.26±0.48
N-acetylated metabolite 1 (1.99, s)	11.9±1.8	2.7×10^{-8}	0.87±0.29	14.0±2.3	1.4×10^{-6}	0.95±0.38	10.6±2.9	1.3×10^{-3}	0.88±0.48		n.s.	
N-acetylated metabolite 2 (2.01, s)	36.4±2.3	$<2.2\times10^{-16}$	1.78±0.33	33.2±3.0	5.3×10^{-15}	1.47±0.40	40.9±3.4	1.6×10^{-14}	2.60±0.62	43.4±4.9	1.4×10^{-15}	1.72±0.51
N-acetylated metabolite 3 (2.03, s)	20.1±2.3	4.0×10^{-12}	1.07±0.30	21.6±2.6	1.9×10^{-9}	1.16±0.39	20.6±4.1	3.5×10^{-5}	1.08±0.49	33.6±4.2	1.1×10^{-10}	1.57±0.50
N-acetylated metabolite 4 (2.05, s)	19.9±1.7	$<2.2\times10^{-16}$	1.42±0.31	22.2±2.2	1.0×10^{-12}	1.55±0.41	17.1±3.2	8.5×10^{-6}	1.25±0.50	19.9±1.7	$<2.2\times10^{-16}$	1.42±0.31
N-acetylated metabolite 5 (2.07, s)	12.5±2.4	1.1×10^{-10}	0.72±0.29	13.1±3.7	3.5×10^{-8}	0.66±0.37	11.5±3.0	8.5×10^{-6}	0.90±0.48	16.7±4.6	4.8×10^{-8}	0.73±0.45
N-acetylated metabolite 6 (2.08, s)	10.7±1.8	1.7×10^{-7}	0.81±0.30	13.8±2.3	2.0×10^{-7}	1.03±0.38		n.s.		16.4±2.6	1.0×10^{-7}	1.34±0.49

Table 5.8 (continued)

Metabolite (δ, multiplicity)	Cancer vs. Control (ALL SAMPLES)			Cancer vs. Control (MALES ONLY)			Cancer vs. Control (FEMALES ONLY)			Cancer vs. Control (AGE-MATCHED)		
	% variation	*p*-value[a]	effect size	% variation	*p*-value[a]	effect size	% variation	*p*-value[a]	effect size	% variation	*p*-value[a]	effect size
Phenylacetylglutamine (2.26, m) [c]	20.6±4.9	2.7×10^{-4}	0.53±0.28	31.6±6.1	6.7×10^{-5}	0.74±0.37		n.s.			n.s.	
Tartrate (4.35, s)	-69.8±21.0	$<2.2\times10^{-16}$	-0.83±0.29	-76.0±29.8	1.1×10^{-14}	-1.11±0.39	-58.1±26.7	9.0×10^{-6}	-0.59±0.47	-73.6±25.5	1.2×10^{-9}	-0.95±0.46
Trigonelline (9.12, s)	-49.8±11.4	2.5×10^{-10}	-0.89±0.29	-51.1±15.7	1.9×10^{-6}	-1.00±0.38	-50.6±17.8	2.3×10^{-6}	-0.80±0.48	-49.7±16.9	7.3×10^{-5}	-0.84±0.46
Unknown (4.30, m)	24.9±2.4	$<2.2\times10^{-16}$	1.29±0.31	28.0±2.6	7.0×10^{-13}	1.56±0.41	23.2±4.6	2.5×10^{-6}	1.10±0.49	24.9±2.4	$<2.2\times10^{-16}$	1.3±0.31
Unknown (5.35, m)	50.6±4.8	1.6×10^{-13}	1.16±0.30	38.5±6.0	3.1×10^{-06}	0.93±0.38	65.4±8.5	1.5×10^{-8}	1.46±0.51	58.5±7.1	5.6×10^{-9}	1.45±0.49
Unknown (7.68, d)	28.3±1.9	$<2.2\times10^{-16}$	1.80±0.33	34.4±2.3	$<2.2\times10^{-16}$	2.16±0.45	23.5±3.1	7.4×10^{-9}	1.57±0.52	31.3±2.9	2.3×10^{-13}	2.12±0.55

[a]Wilcoxon rank sum test $p<1.4\times10^{-3}$ (Bonferroni-corrected). Metabolites found to be affected by [b]gender- or [c]age-related bias.

5.5.2 Age-related metabolic features in urine

The significant difference between cancer and control subjects (median ages of 63 and 45 years old, respectively) can potentially bias the results of group discrimination. The extent to which age was reflected on the urinary profile was firstly assessed by OPLS1 regression analysis using the spectra from controls and their age as the Y-matrix (age range 22-59). The resulting scores scatter plot suggested a relationship with age, with older subjects being located towards positive LV1 (Figure 5.10a). Despite the low predictive power obtained by seven-fold cross validation (Q^2 0.13), the corresponding loadings (Figure 5.10b) were examined and suggested the following age-related metabolic features: increased excretion of citrate, dimethylamine, trimethylamine-*N*-oxide, hippurate and phenylacetylglutamine in older subjects, and increased levels of *N*-acetylated metabolites, creatinine and histidine in younger subjects.

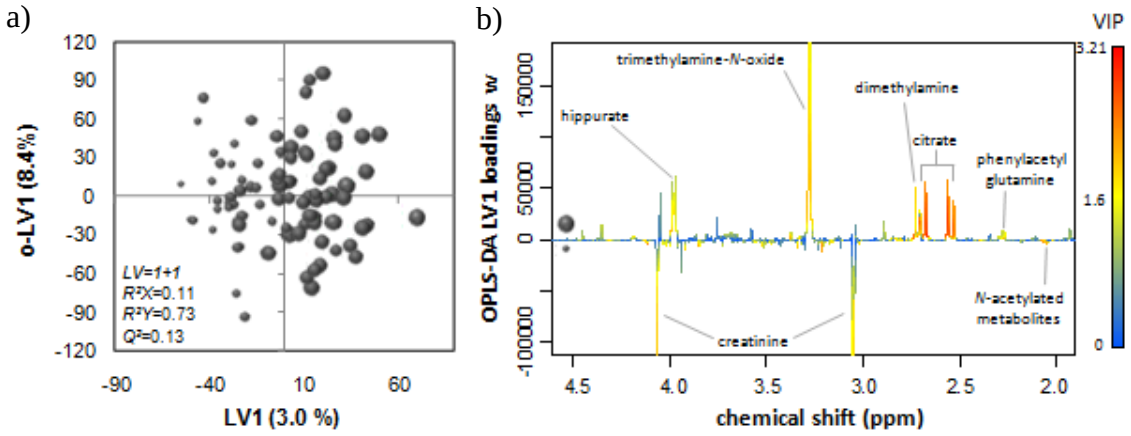


Figure 5.10 OPLS regression of the urine ^{1}H NMR spectra of healthy controls (n 90) and the subjects' age: a) scores scatter plot of the first two latent variables, where the size of the circle is proportional to age, b) LV1 loadings weights coloured as a function of VIP. The parameters shown on the scores plot (*LV*: latent variable, R^2X: variation explained by the X matrix, R^2Y: variation explained by the Y matrix, Q^2: predictive power) derive from default 7-fold cross validation.

The increased excretion of trimethylamine-*N*-oxide, dimethylamine and citrate in older subjects has been already reported by Psihogios and co-workers (Psihogios et al. 2008). They suggested a slight dysfunction of the osmotic role of these metabolites in renal medullary tissue as the main reason for this fact. Also, in accordance with what was observed in this work, another study involving a Taiwanese and American population, reported a positive correlation between the levels of phenylacetylglutamine and age (Swann et al. 2013). The decline of creatinine levels in older subjects (Slupsky et al. 2007) has been associated with decreasing muscle mass and decline in glomerular filtration rate (Salek et al. 2007). Concerning hippurate levels, an opposite variation (decrease with age) was published by Psihogios and no reference to histidine levels was found. Interestingly, blood plasma histidine has also been highlighted in the previous chapter to be dependent on age, having shown increased levels in younger subjects (as seen for urine).

By comparing age- and cancer-related variables (Table 5.9) and given that the cancer group comprised much older subjects than the control group, it becomes apparent that a few of the latter variables may potentially be biased by the age difference between groups. This is the case of citrate and phenylacetylglutamine, simultaneously increased with age and in cancer, and of histidine, simultaneously decreased in younger subjects and in controls.

Table 5.9 Age-related metabolic variations in urine and their comparison with cancer-related variations.

Age-related metabolites	Variation with increasing age	Variation cancer vs. control	Observations
N-acetylated metabolites	↓	↑	Opposite variation
Citrate	↑	↑	Possible bias[a]
Creatinine	↓	↑	Opposite variation
Dimethylamine	↑	-	
Hippurate	↑	↓	Opposite variation
Histidine	↓	↓	Possible bias[a]
Phenylacetylglutamine	↑	↑	Possible bias[a]
Trimethylamine-*N*-oxide	↑	-	

[a] Metabolites increased/decreased in both older subjects and patients may potentially bias cancer vs. control discrimination as the median age of the cancer group is significantly higher than that of the control group.

In order to further investigate this possibility, a subgroup of subjects with a smaller age difference between controls (n 42, median age 50 years-old, 20M/22F) and patients (n 38, median age 56 years-old, 22M/16F) has been considered for multivariate analysis. Although absolute matching was not possible, the influence of age is expected to be significantly reduced in this subgroup of samples (hereby designated as age-matched). Notably, a model with high predictive power (Q^2 0.70) and classification rate (90.3%) could still be obtained (Table 5.6), thus allowing age to be excluded as a strong confounding factor. Concerning the comparison of cancer-related variables highlighted in the global and age-matched models (Figure 5.11), there are a few differences to note. Citrate, histidine and phenylacetylglutamine did not have VIP>1 or statistically significant differences between control and cancer age-matched groups, thus confirming that their importance as cancer markers in the global model could be biased by age mismatching. Moreover, spectral integration also demonstrated the loss of significance for other metabolites, previously mentioned as important for group discrimination, namely creatine, *p*-cresol sulphate, 4-deoxythreonate, fatty acids, hippurate and *N*-acetylated metabolite δ1.99 (Table 5.8). This result suggested that the metabolic response to cancer may differ with age.

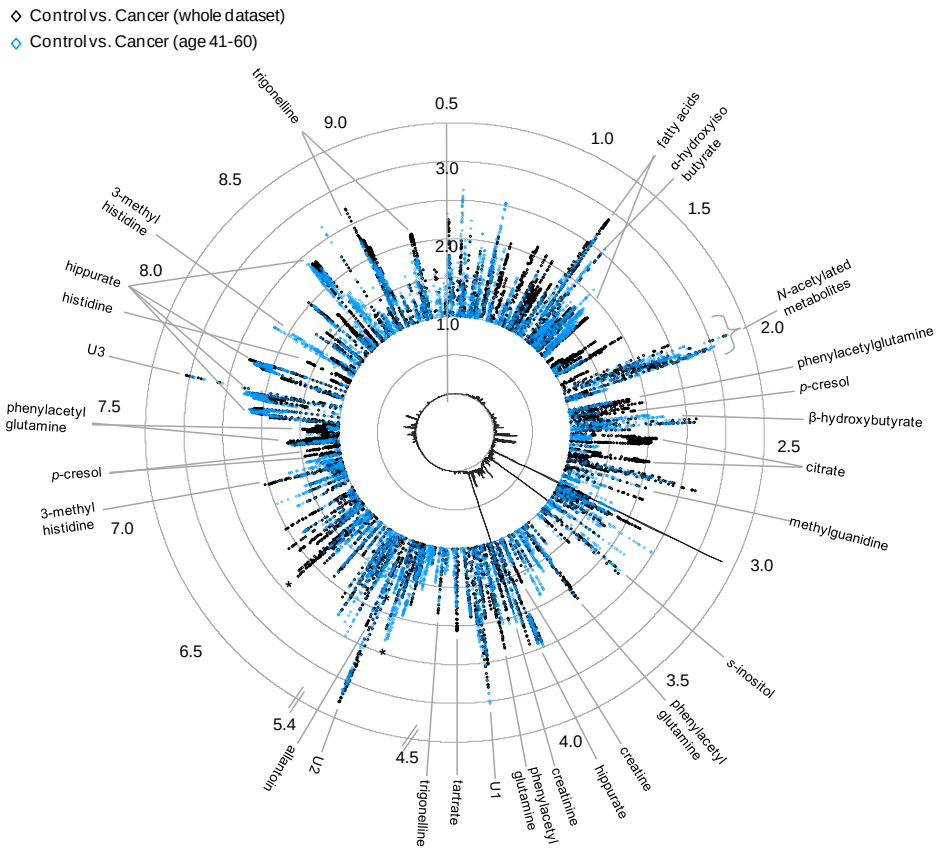


Figure 5.11 VIP-wheel representation of variables found to have VIP>1 in the urine OPLS-DA models built with: the whole dataset (black), a subset with improved age-matching between controls and patients (blue).

5.5.3 Possible impact of smoking habits and other potential confounders

For assessing the possible influence of smoking habits on the urinary metabolic composition, control subjects were divided into smokers (n 28) and never-smokes (n 43) and their NMR metabolic profiles compared by means of MVA. MCCV of the PLS-DA model resulted in a moderate classification rate (74.1%) and relatively low predictive power (median Q^2 0.38) (Table 5.6). This is in accordance with a previous urine NMR metabolomics study failing to accurately classify subjects according to their smoking habits (Psihogios et al. 2008). The corresponding loadings also reflected the low impact of smoke on urine spectra, as only a few variables showed VIP>1 (not shown). These included the resonances of trigonelline (increased in smokers' urine) and several unknown signals, some of which being correlated in STOCSY plots (not shown): doublet at δ 2.18 correlated with singlet at δ 9.05, and singlet at δ 4.41 correlated with doublet at δ 8.79 (increased in smokers' urine), together with singlets at δ 3.16 and δ 3.65 ppm

(decreased in smokers urine). The possible assignment of these signals to common nicotine urinary metabolites (nicotine and cotinine derivatives) has been attempted with no success, their identification remaining unknown. In regard to trigonelline, given that the cancer group is assumed to be composed mainly by smoke-exposed subjects (based on the histological characteristics of their lungs), the opposite variations found for this metabolite in smokers vs. non-smokers and cancer vs. control comparisons enables its possible bias to be discarded. The role of smoking habits as a possible confounding factor in cancer vs. control discrimination was further assessed by considering multivariate modelling of each control subgroup (never-smokers and smokers) against the cancer group. Both models maintained high predictive power (Q^2 values 0.6-0.7) and classification rate (>90%). Moreover, the loadings profile resulting from these models (not shown) were identical to those obtained for the global model, thus confirming the small influence of smoking habits on cancer vs. control discrimination.

With respect to other possible confounding factors, diet is certainly one of the most relevant, as it is well known that it can modulate human urinary metabolic composition and account for the high inter-individual variability observed in urine spectra (Stella et al. 2006; Walsh et al. 2006; Winnike et al. 2009; Rasmussen et al. 2011). However, in the majority of metabolomic studies there is no dietary restriction, likely to better reflect a possible clinical scenario, so that the interpretation of disease biomarkers in urine should take the variability introduced by diet into consideration. In this work, dietary influence and impact of time of day or circadian rhythms were minimized by collecting all samples in the morning, after overnight fasting, as suggested in the literature (Slupsky et al. 2007). Still, possible relationships of cancer-related metabolic variations with diet and/or gut microflora may be present and will be addressed in section 5.8. Body mass index (BMI) has also been reported to be reflected on urine composition, namely on the levels of citrate, lactate, dimethylamine and glycine (Kochhar et al. 2006). Amongst these metabolites, only citrate has been putatively identified in this work as being cancer-related (as well as age-related). Therefore, although BMI data was not available for this study, based on the literature, it is unlikely that it might have a strong influence on the results obtained.

5.6 Reassessment of cancer vs. control discrimination after excluding possibly biased variables

Consideration of gender-, age- and smoke-related metabolite variations (described in the sections above) led to the identification of citrate, creatine, creatinine, histidine and phenylacetylglutamine as possibly biased variables. Thus, reassessment of cancer vs. control discrimination after exclusion of those metabolites has been performed, either using full-spectral resolution or the relevant integrals for class discrimination. Figure 5.12a-c shows the MVA results obtained for the comparison of controls and lung cancer patients, when using NMR full-resolution data, after excluding biased variables.

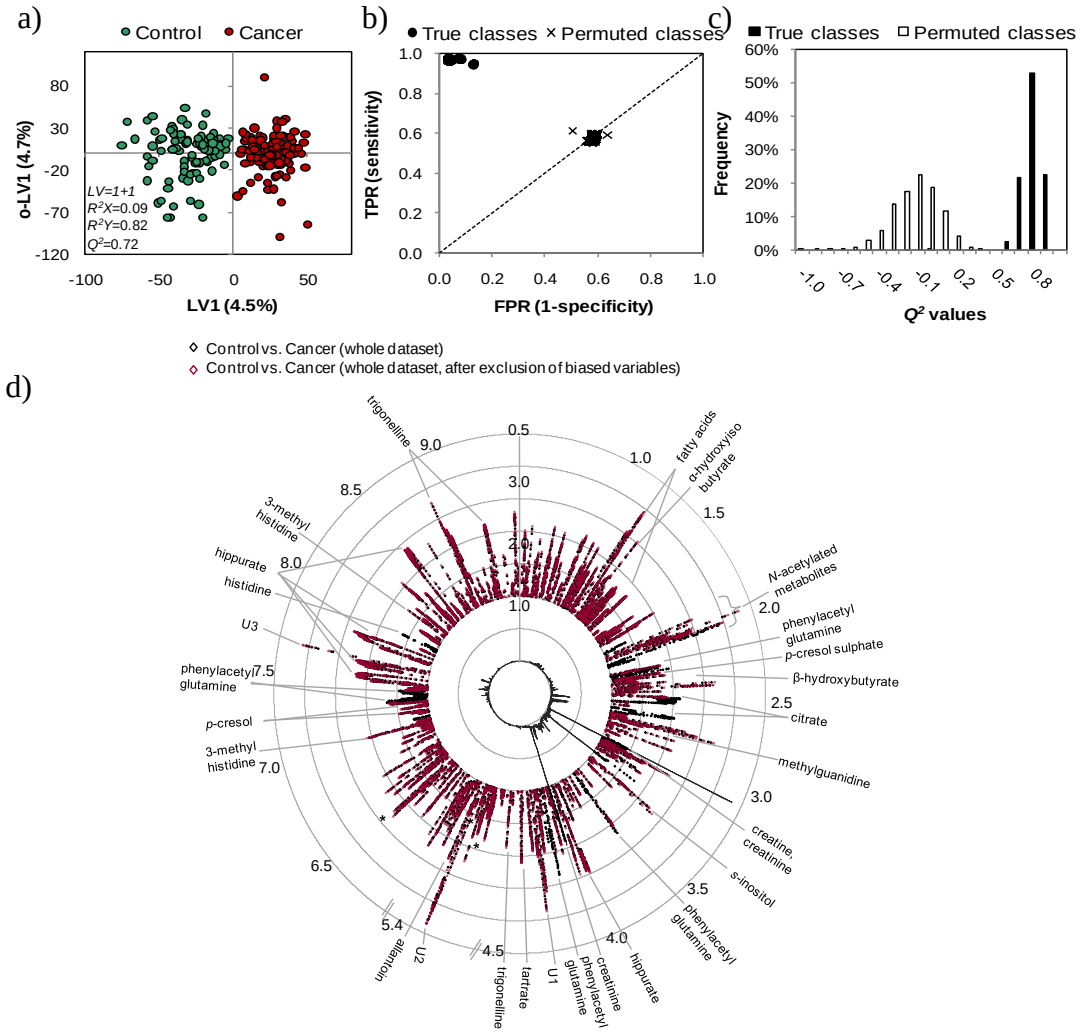


Figure 5.12 MVA applied to the standard ^{1}H NMR spectra of urine from controls (n 91) and cancer patients (n 109), after exclusion of gender- and age-biased variables: a) OPLS-DA scores scatter plot, b) ROC space (TPR: true positive rate, FPR: false positive rate) and c) Q^2 histogram obtained by MCCV and permutation testing (500 iterations) of the PLS-DA models. d) VIP-wheel representation of variables found to have VIP>1 in the OPLS-DA models built with: the whole spectral range (black) or after exclusion of biased variables (brown). The parameters shown on the scores plot (*LV*: latent variable, R^2X: variation explained by the X matrix, R^2Y: variation explained by the Y matrix, Q^2: predictive power) derive from default 7-fold cross validation.

The classification results were very similar to the model obtained previously (without rejecting any variables) (Table 5.10), indicating the low impact of the excluded variables. Also, as seen in the VIP-wheel representation (Figure 5.12d), apart from the difference in excluded variables, there was almost perfect overlap between the important variables highlighted in the two models.

Table 5.10 Prediction results obtained by MCCV (500 iterations) of urine PLS-DA models assessing the discrimination between lung cancer patients and healthy controls: comparison with models reassessed after exclusion of gender- and age-biased variables.

PLS-DA models	Median Q^2	Sensitivity (%)	Specificity (%)	Classification rate (%)
Standard 1D	0.76	97.3	96.2	96.8
Standard 1D (after exclusion of biased variables)	0.75	96.9	95.6	96.3
24 Integrals	0.72	97.2	94.2	95.8
19 Integrals (after exclusion of biased variables)	0.71	94.8	96.2	95.4

A similar evaluation was performed using the integrals found to be relevant in cancer vs. control discrimination. So, when excluding citrate, creatine, creatinine, histidine and phenylacetylglutamine from the group of twenty four important integrals, MCCV afforded comparable quality parameters to those obtained for the model comprising all the integrals (Table 5.10 and Figure 5.13). Moreover, compared to the results obtained for the model built with full-resolution spectra, MVA of the nineteen relevant integrals provided equally good classification rate and predictive ability, suggesting these metabolites to be representative of the urinary malignant signature.

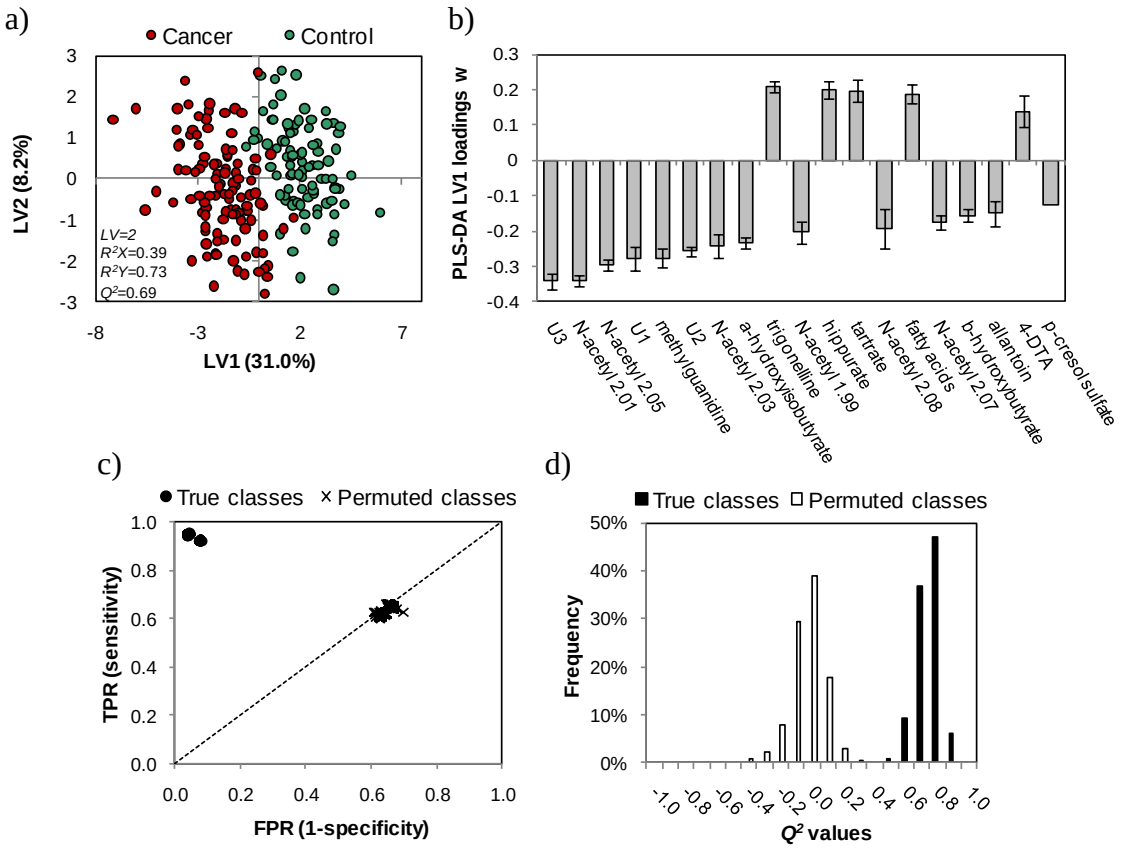


Figure 5.13 PLS-DA applied to 19 integrals (after exclusion of gender- and age-biased variables) measured in the standard 1D 1H NMR spectra of urine from controls (n 91) and cancer patients (n 109): a) scores scatter plot and b) LV1 loadings weights. The parameters shown on the scores plot (*LV*: latent variable, R^2X: variation explained by the X matrix, R^2Y: variation explained by the Y matrix, Q^2: predictive power) derive from default 7-fold cross validation. c) ROC space (TPR: true positive rate, FPR: false positive rate) and d) Q^2 histogram obtained by MCCV and permutation testing (500 iterations).

5.7 Preliminary external validation of urine-based classification models

Although a truly independent sample set was not available in this study, the ability of the urine profile to predict class membership of new samples was tested by rebuilding the classification model with a smaller group of samples (training set: control n 71, cancer n 89) and using this model to predict the class of the remaining samples (prediction set: control n 20, cancer n 20). The samples in the prediction set were those collected and analysed more recently. PLS-DA models were built using either the full spectral range (excluding the biased variables identified in the previous section) or the areas of the 19 metabolites found to be more relevant in cancer vs. control discrimination. The respective scores scatter plots are shown in Figure 5.14a, b and the classification results are presented in Table 5.11. Both models predicted correctly 18 out of 20 controls (90%

specificity), whereas the integrals-based model showed superior sensitivity (all cancer samples were correctly classified versus one sample misclassified by the full resolution model). These results are quite promising as they show the potential usefulness of urine NMR metabolomics as an adjunct tool in lung cancer detection and screening. Further validation of this approach should therefore entail the analysis of an enlarged sample set (moving from hundreds to thousands of samples), and the independent measurement of the putative metabolite markers by other quantitative methods. Moreover, it would be most interesting to assess the specificity of the urinary signature revealed in comparison with other diseases, especially other lung diseases.

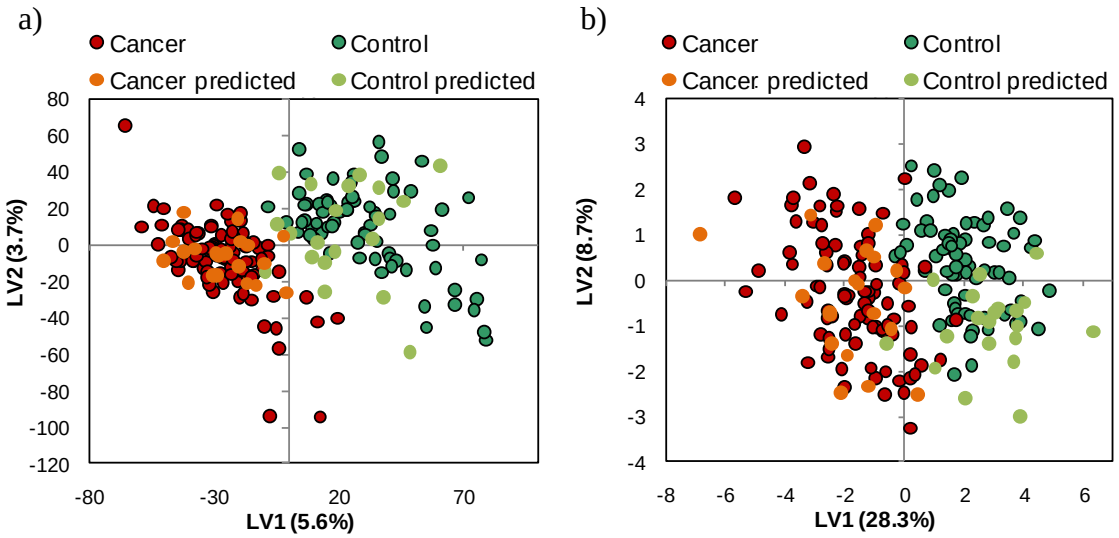


Figure 5.14 Scores scatter plots obtained by PLS-DA of a) full resolution ^{1}H NMR urine spectra (LV=2, R^2X=0.09, R^2Y=0.83, Q^2=0.70) and b) 19 signal integrals (LV=2, R^2X=0.37, R^2Y=0.74, Q^2=0.70), using in the calibration set 71 samples from controls and 89 samples from patients. The scores corresponding to class prediction of additional 40 samples (20 controls and 20 patients) are superimposed in lighter colour.

Table 5.11 Confusion matrix showing the prediction results for an independent test set (20 controls and 20 patients) obtained by PLS-DA modelling of a calibration set (71 controls and 89 patients), using either the full resolution urine spectra (after removal of biased variables) or 19 selected signal integrals. Shaded boxes show the number of samples correctly classified.

			Full resolution spectra		19 Integrals				
			True class		True class				
			Cancer	Control	Cancer	Control			
Full resolution spectra	Predicted class	Cancer	19	2	20	2	Cancer	Predicted class	19 Integrals
		Control	1	18	0	18	Control		

5.8 Proposed biochemical interpretation of cancer-related metabolic variations in urine

According to the results presented in the previous sections, the putative urinary metabolic signature of lung cancer comprised the main following metabolic features: increase levels of β-hydroxybutyrate, α-hydroxyisobutyrate, allantoin, methylguanidine, *p*-cresol sulphate, six *N*-acetylated metabolites and three unknowns, together with decreased levels of fatty acids, trigonelline, tartrate, hippurate and 4-deoxythreonate.

In agreement with the plasma results, β-hydroxybutyrate has been found increased in the urine of cancer patients, particularly AdC patients. As previously mentioned (subchapter 4.7), the presence of this ketone body in biofluids may indicate increased fatty acid oxidation. The observed decrease in urinary levels of fatty acids further supports this hypothesis, corroborating the important role of this pathway in cancer metabolism, as recently brought to light by other studies (Carracedo et al. 2013). Interestingly, this metabolite has also been associated with muscle loss in patients with lung and colon cancers (Eisner et al. 2011). Another hydroxy acid found to be increased in the patients' urine was α-hydroxyisobutyrate, which has also been associated with muscle loss (Eisner et al. 2011), and may reflect the abnormal organic acid excretion resulting from metabolic decompensation (Kumps et al. 2002).

Allantoin is a diureide of glyoxylic acid resulting from the free radical-induced oxidation of uric acid (the terminal product of purine metabolism) (Wishart et al. 2007). As its urinary levels have been shown to reflect the systemic load of reactive oxygen species (ROS), independently of uric acid levels, allantoin has been has been suggested as a potential biomarker for oxidative stress (Tolun et al. 2010). Increased levels of allantoin have been associated with several diseases, including diabetes, inflammatory and autoimmune conditions, cardiovascular and pulmonary diseases (Tolun et al. 2010). In this work, allantoin was found elevated in the urine of lung cancer patients, especially SqCC. Although no previous accounts on the variation of this metabolite in lung cancer were found, increased urinary allantoin has also been reported in ovarian cancer patients (Slupsky et al. 2010). Moreover, in a mouse model of *S. aureus* lung infection, urinary allantoin was found to increase upon treatment of infected animals (Slupsky et al. 2009).

Methylguanidine is a guanidine compound deriving from protein catabolism (Wishart et al. 2007) or from the microbial conversion of creatinine (Wyss and Kaddurah-

Daouk 2000). Increased plasmatic levels of this metabolite have been associated with renal failure (Wishart et al. 2007), while its possible anti-inflammatory activity has also been demonstrated in a mice model of septic shock (Marzocco et al. 2004). Its possible association to cancer has, however, not been documented, except for a study of oesophageal cancer where, similarly to our results, methylguanidine was found elevated in the urine of patients (Hasim et al. 2012).

Increased *p*-cresol sulphate was another feature of the patients' urinary profile, particularly of AdC patients. This metabolite likely derives from gut microbial-host co-metabolism and its abnormal systemic levels have been linked to multiple sclerosis, cardiovascular disease and oxidative injury (Wishart et al. 2007). In the context of cancer, it has been found increased in the urine of colorectal patients (Qiu et al. 2010), although an opposite variation was reported in a subsequent study (Cheng et al. 2012).

A number of signals identified as *N*-acetylated metabolites were all found increased in the urine of lung cancer patients, compared to healthy controls. The increased urinary excretion of *N*-acetylated compounds, including *N*-acetylated amino acids and *N*-acetylneuraminic acid (Neu5Ac), has been linked to inborn errors of metabolism (Engelke et al. 2004). In regard to cancer studies, urinary *N*-acetylaspartate was found to be significantly increased in preoperative colorectal cancer patients (Qiu et al. 2010), while Neu5Ac was recently reported to be elevated in both tissues and urine of non-small cell lung cancer patients (Mathe et al. 2014). The authors highlighted the role of Neu5Ac in cell signalling, binding and transportation of positively charged molecules, in particular it's protecting role of malignant cells from cellular defence systems. Therefore, having these findings into account, an important following step in this work would be to identify the specific *N*-acetylated metabolites contributing for the urinary signature of lung cancer, in order to further understand their role in cancer metabolism and further assess their diagnostic potential.

In regard to metabolites showing lower levels in the urine of cancer patients relatively to controls, most of them seem to share a possible relationship with diet or gut microbial activity. Trigonelline is a dietary betaine present in several foodstuffs (e.g., coffee), and a product of niacin (vitamin B3) metabolism (Wishart et al. 2007). In a study addressing lung function markers in patients with chronic obstructive pulmonary disease (COPD), increased urinary levels of trigonelline were linked to improved lung function (McClay et al. 2010), whereas decreased levels of this metabolite were observed in the

urine of mice infected with *S. pneumonia* (Slupsky et al. 2009). In cancer studies, urinary trigonelline was found to be decreased in patients with pancreatic ductal adenocarcinoma (Napoli et al. 2012), bladder cancer (Chen et al. 2012) and ovarian cancer (Slupsky et al. 2010), similarly to the variation hereby observed. Moreover, trigonelline has been highlighted as one of the metabolites correlating with muscle loss in colorectal and lung cancers at advanced stage (Eisner et al. 2011). Tartrate is another diet-related metabolite, associated with the consumption of wine and fruits (Heinzmann et al. 2012) and its possible association with cancer has not been documented.

Hippurate, an acyl glycine resulting from the conjugation of benzoic acid with glycine, is a urinary metabolite arising from several sources, including diet (consumption of phenolic-containing products like tea, wine, fruits), exposure to solvents (e.g., toluene) or oxidative stressors (e.g., tobacco smoke) and intestinal microfloral activity. Although one UPLC-MS study has reported increased hippurate in the urine of lung cancer patients compared to controls (Yang et al. 2010), thus contradicting our results, decreased urinary hippurate levels have often been reported across several cancer types, namely hepatocellular cancer (Shariff et al. 2011), oesophageal cancer (Hasim et al. 2012), oral cancer (Xie et al. 2012), osteosarcoma (Zhang et al. 2010), bladder cancer (Van et al. 2011), pancreatic ductal adenocarcinoma (Napoli et al. 2012) and ovarian and breast cancers (Slupsky et al. 2010).

In regard to decreased levels of 4-dexoythreonate (4-DTA), no interpretation could be proposed at this point. 4-DTA is a common urinary metabolite and has been proposed to derive from *L*-threonine (Kassel et al. 1986). In early works it has been associated with *diabetes mellitus* type 1 (Kassel et al. 1986) and uraemia (Bultitude and Newham 1975), but no association to cancer was found in the literature.

As a final remark, it is interesting to note that, while in the case of tissues and plasma there were a number of metabolic features pointing directly to pathways known to play central roles in cancer metabolism, urinary profiling highlighted a number of novel variations which are difficult to explain in the context of current knowledge. While this may be somehow frustrating, it also represents a great stimulus to pursue further work towards a deeper understanding of urinary changes in cancer and it underlines the importance of looking into different, complementary metabolic windows.

6 PRELIMINARY UPLC-MS METABOLOMIC STUDY OF LUNG CANCER URINARY ALTERATIONS

6.1 UPLC-MS profile of urine

The analysis of urine by ultra-performance liquid chromatography coupled to mass spectrometry (UPLC-MS) allowed four different datasets to be obtained, as each sample was separately run in two different chromatographic columns (reverse-phase high strength silica – HSS and hydrophilic interaction chromatography – HILIC) and detected using two ionisation modes (electrospray positive – ESI+ and negative – ESI-). Figure 6.1 illustrates representative total ion chromatograms (TIC) corresponding to the four different datasets acquired, demonstrating the ability of this approach for obtaining complementary metabolic information depending on the column and ionisation mode chosen.

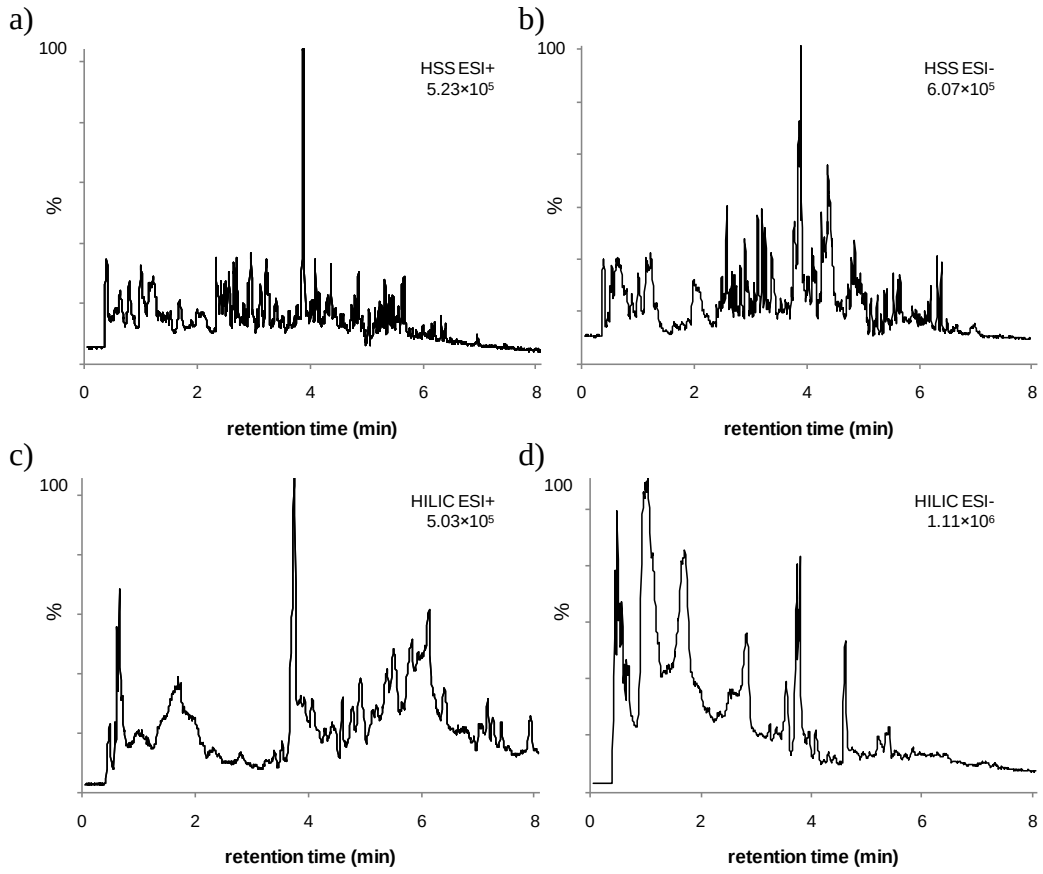


Figure 6.1 Representative UPLC-MS total ion chromatograms (TIC) of urine from a lung cancer patient, acquired in two chromatographic columns and two ionisation modes: a) HSS ESI+, b) HSS ESI-, c) HILIC ESI+ and d) HILIC ESI-.

The quality of UPLC-MS data was evaluated by injecting a quality control (QC) sample, composed of aliquots from all urines used in the run, at regular intervals (every ten samples), from which stability/variability of the run could be assessed (Want et al. 2010). Figure 6.2 shows the scores scatter plot resulting from applying PCA to the study samples (n 98) and to the in-run QCs (n 10), after applying different scaling procedures to the data (to minimize the dominance of higher intensity signals over less intense ones).

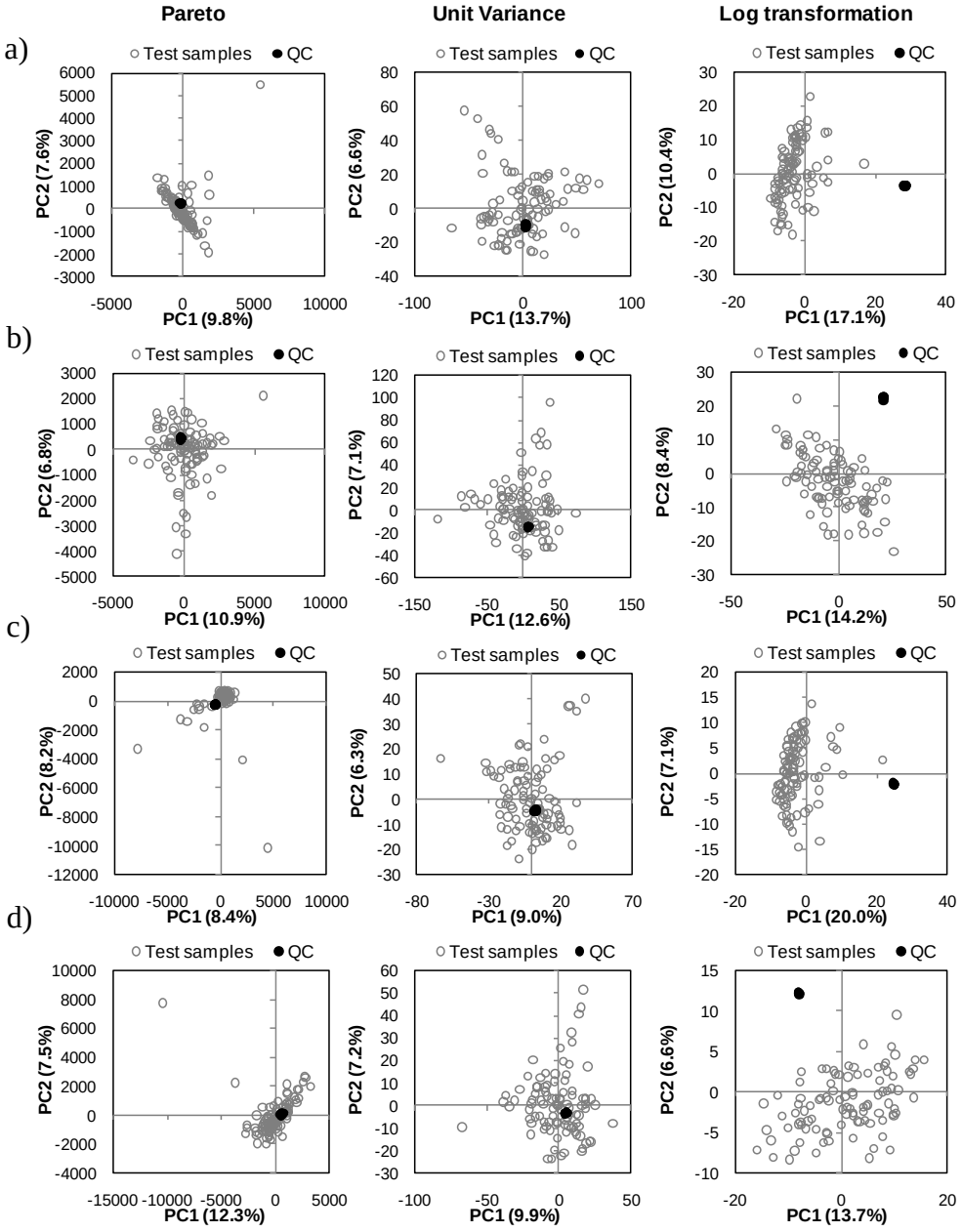


Figure 6.2 PCA scores scatter plots of urine UPLC-MS datasets of QC (●, n 10) and study (○, n 98) samples using Pareto scaling (left), unit variance scaling (centre) and logarithm transformation (right): a) HSS ESI+, b) HSS ESI-, c) HILIC ESI+, d) HILIC ESI-.

The fact that QC samples were highly overlapped (no dispersion in the scores scatter plot) indicated good stability throughout the four chromatographic runs, which is crucial for high quality data to be recorded and further used in multivariate analysis.

6.2 Potential of urine UPLC-MS profile to discriminate between patients and control subjects

After UPLC-MS data pre-treatment (described in section 2.4.3), PCA and PLS-DA were applied to all twelve datasets (two columns × two ionization modes × three scaling/transformation types) composed by the control group (n 50, 25 males, 25 females, average and median age 49, age range 38-59) and the patients group (n 48, 37 males, 11 females, average and median age 61, age range 30-84). Further details on patients' histological and stage classifications can be found in Table 2.1. Figure 6.3 shows the PLS-DA scores scatter plots obtained for all datasets. Similarly to what was obtained for ^{1}H NMR data of the same sample set (not shown), PLS-DA of UPLC-MS datasets (with different scaling/transformation applied) allowed control and cancer urines to be well separated along LV1 (Figure 6.3). Model validation was performed by means of MCCV and permutation testing (500 iterations) and the corresponding prediction results are shown in Table 6.1. In the ROC space plots, models with permuted classes fell along the diagonal line (no discrimination), whereas models with true classes assigned were located in the upper left corner (Figure 6.4), showing high sensitivity and specificity. Regarding Q^2, histograms showed little overlap between models with true and permuted classes, demonstrating once more the robustness of the models built (Figure 6.5). Overall, classification rates were higher than 90%, sensitivity and specificity ranged between 88-97% and 89-100%, respectively, and median Q^2 ranged between 0.66-0.80. No column type, ionisation mode or scaling/transformation procedure performed consistently better than the others, demonstrating the usefulness of acquiring data in different experimental setups and of applying different scaling/transformation procedures in data pre-treatment.

Although showing promise and corroborating the potential diagnostic value of urinary metabolic profiles, also evidenced by NMR metabolomics, this study was limited by imperfect mismatching of control and cancer groups in terms of gender and age. Therefore, a subsequent stage of this work should include careful assessment of their

possible confounding influence (as performed for NMR data), as well as of the possible dependency of UPLC-MS metabolic profiles on cancer histological type and stage.

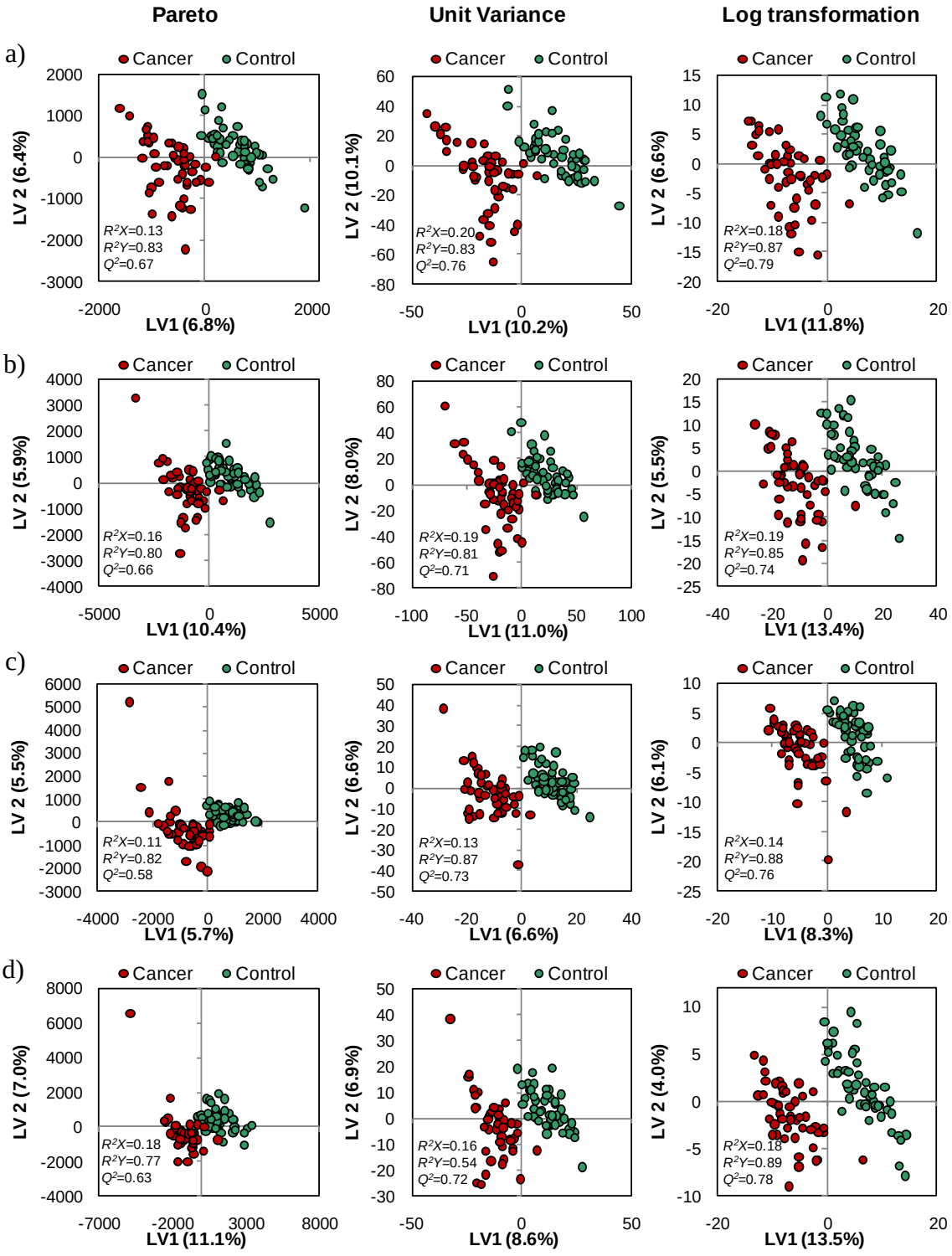


Figure 6.3 PLS-DA scores scatter plots of urine UPLC-MS datasets from controls (n 50) and cancer patients (n 48) using Pareto scaling (left), unit variance scaling (centre) and logarithm transformation (right): a) HSS ESI+, b) HSS ESI-, c) HILIC ESI+, d) HILIC ESI-.

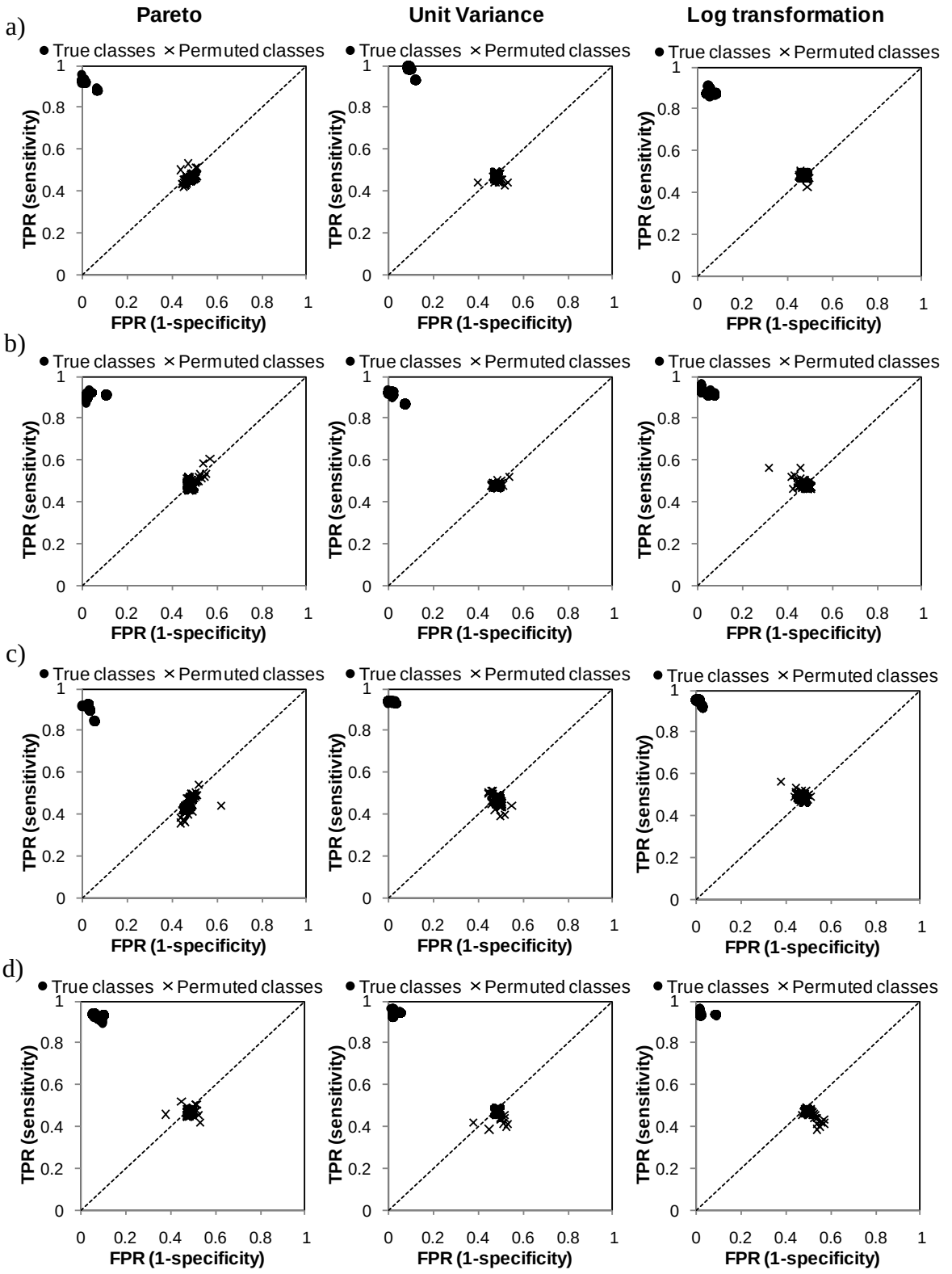


Figure 6.4 ROC space plots (TPR: true positive rate; FPR: false positive rate) obtained by MCCV and permutation testing (500 iterations) of the PLS-DA models built with UPLC-MS urine datasets from controls (n 50) and cancer patients (n 48) using Pareto scaling (left), unit variance scaling (centre) and logarithm transformation (right): a) HSS ESI+, b) HSS ESI-, c) HILIC ESI+, d) HILIC ESI-.

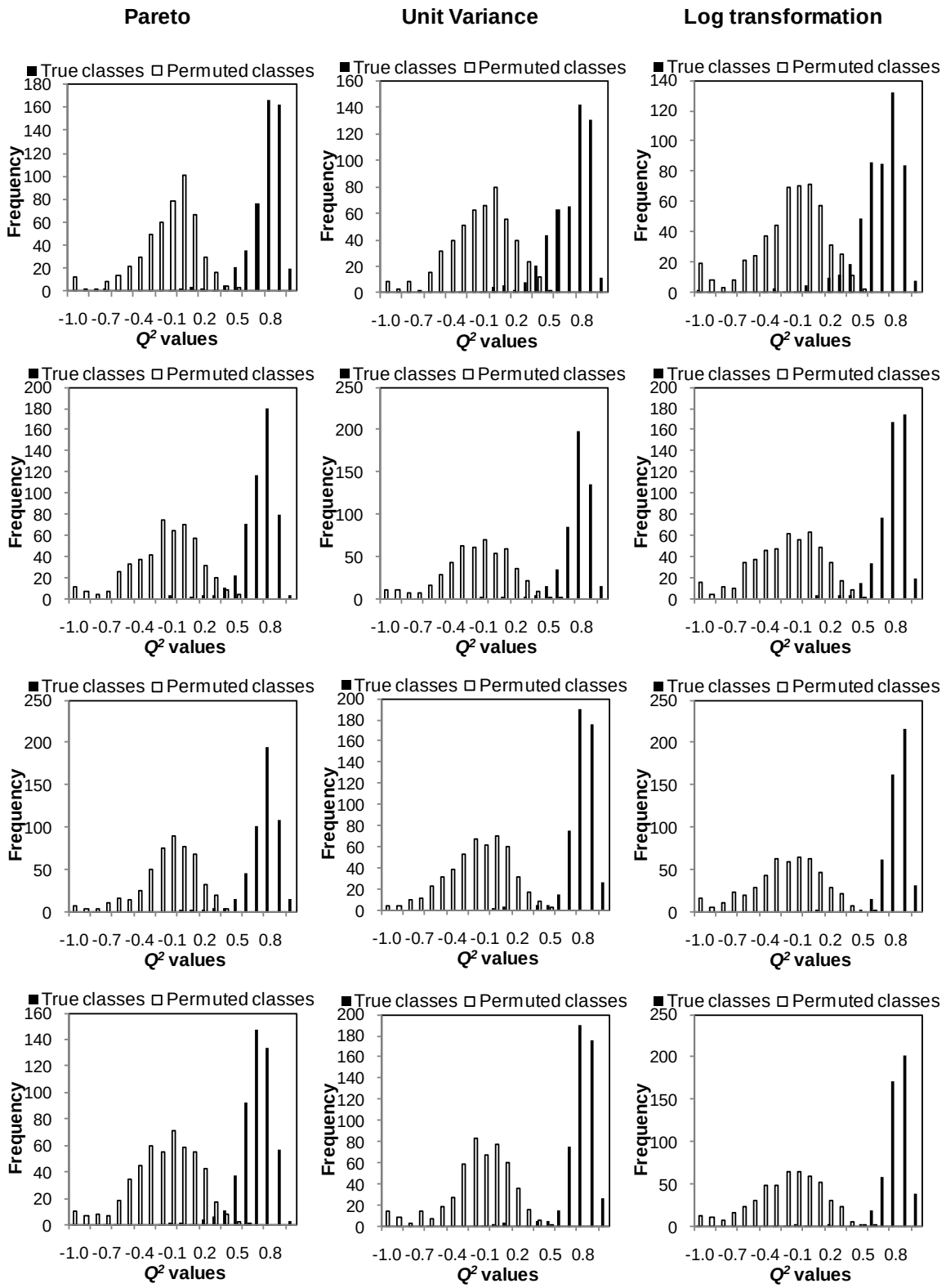


Figure 6.5 Q^2 histograms obtained by MCCV and permutation testing (500 iterations) of the PLS-DA models built with UPLC-MS urine datasets from controls (n 50) and cancer patients (n 48) using Pareto scaling (left), unit variance scaling (centre) and logarithm transformation (right): a) HSS ESI+, b) HSS ESI-, c) HILIC ESI+, d) HILIC ESI-.

Table 6.1 Prediction results obtained by MCCV (500 iterations) of urine UPLC-MS PLS-DA models.

PLS-DA classes [**n samples**, n M/n F, median age]	Type of scaling	Median Q^2	Sensitivity (%)	Specificity (%)	Classification rate (%)
			HSS ESI+		
	UV	0.73	98.6	91.2	94.8
	Par	0.76	92.5	100.0	96.3
	Log	0.66	87.9	94.3	91.2
			HSS ESI-		
	UV	0.74	92.4	98.0	95.2
	Par	0.71	91.2	89.3	90.2
	Log	0.77	92.6	95.3	95.3
			HILIC ESI+		
Control vs. Cancer [**50**, 25/25, 49] vs. [**48**, 37/11, 61]	UV	0.78	94.2	97.9	96.1
	Par	0.75	92.4	97.2	94.9
	Log	0.80	95.2	100.0	97.6
			HILIC ESI-		
	UV	0.78	93.6	97.3	95.5
	Par	0.66	90.7	90.7	90.7
	Log	0.80	95.3	98.0	96.7
			^{1}H NMR		
	UV	0.66	90.1	91.4	90.7

6.3 Selection and tentative identification of UPLC-MS features relevant to class discrimination

In order to extract the features with highest relevance for discriminating between lung cancer patients and controls, a few selection steps were applied, as exemplified next for the HSS ESI+ Pareto scaling dataset. Firstly, in the S-plot, which depicts the covariance and correlation between metabolites, the features with correlation |p(corr)[1]|>0.6 and covariance p[1]>0.02 were selected (Figure 6.6a). Secondly, the covariance ratio plot was used to select those features with lower jack-knife standard error of covariance across the cross validation rounds ($p[1]_{cvSE}$). The threshold used for variable selection corresponds to a standard error of 25%, so features with $|p[1]|/p[1]_{cvSE}<4$ were excluded (Figure 6.6b). Finally, the statistical difference between classes of selected features ($p<0.01$) was evaluated by univariate analysis, namely by the Wilcoxon rank sum test. Table 6.2 shows the number of selected variables for all the data analysed across each selection step. In total, over 600 MS features were selected.

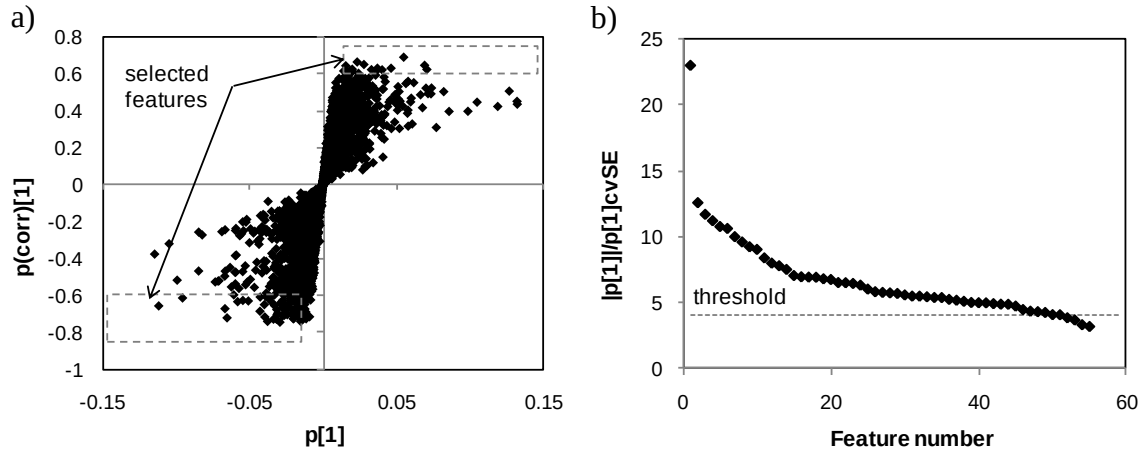


Figure 6.6 Selection of relevant MS features for class discrimination in the urine PLS-DA model obtained for cancer vs. control (HSS ESI+; Pareto scaling). a) S-plot (dashed boxes represent selected features with |p(corr)[1]>0.6 and |p[1]>0.02) and b) covariance ratio plot (dashed line indicates the threshold >4 of covariance ratio above which features were selected).

Table 6.2 Number of relevant variables across each selection step.

	Scaling	**Start number**	**|p(corr)|≥0.6 & |p[1]|≥0.02**	**$\|p[1]\|/p[1]_{cvSE}$ ≥4**	**Significant *p*-value[a]**
HSS ESI+	UV	4434	245	236	50
	Par		113	55	48
	Log		197	196	51
HSS ESI-	UV	7312	453	436	51
	Par		78	75	51
	Log		319	319	51
HILIC ESI+	UV	2628	30	23	23
	Par		6	5	5
	Log		70	70	65
HILIC ESI-	UV	2412	38	38	36
	Par		26	25	24
	Log		162	162	159

[a] Wilcoxon rank sum test.

Figure 6.7 shows the variables selected for each dataset, where the different scaling/transformation are represented by different colours. It is clear in these plots that, although a few variables overlapped, different variables were selected according to the scaling/transformation used. This result highlights the importance of using different processing approaches to analyse this type of data, as they give complementary information.

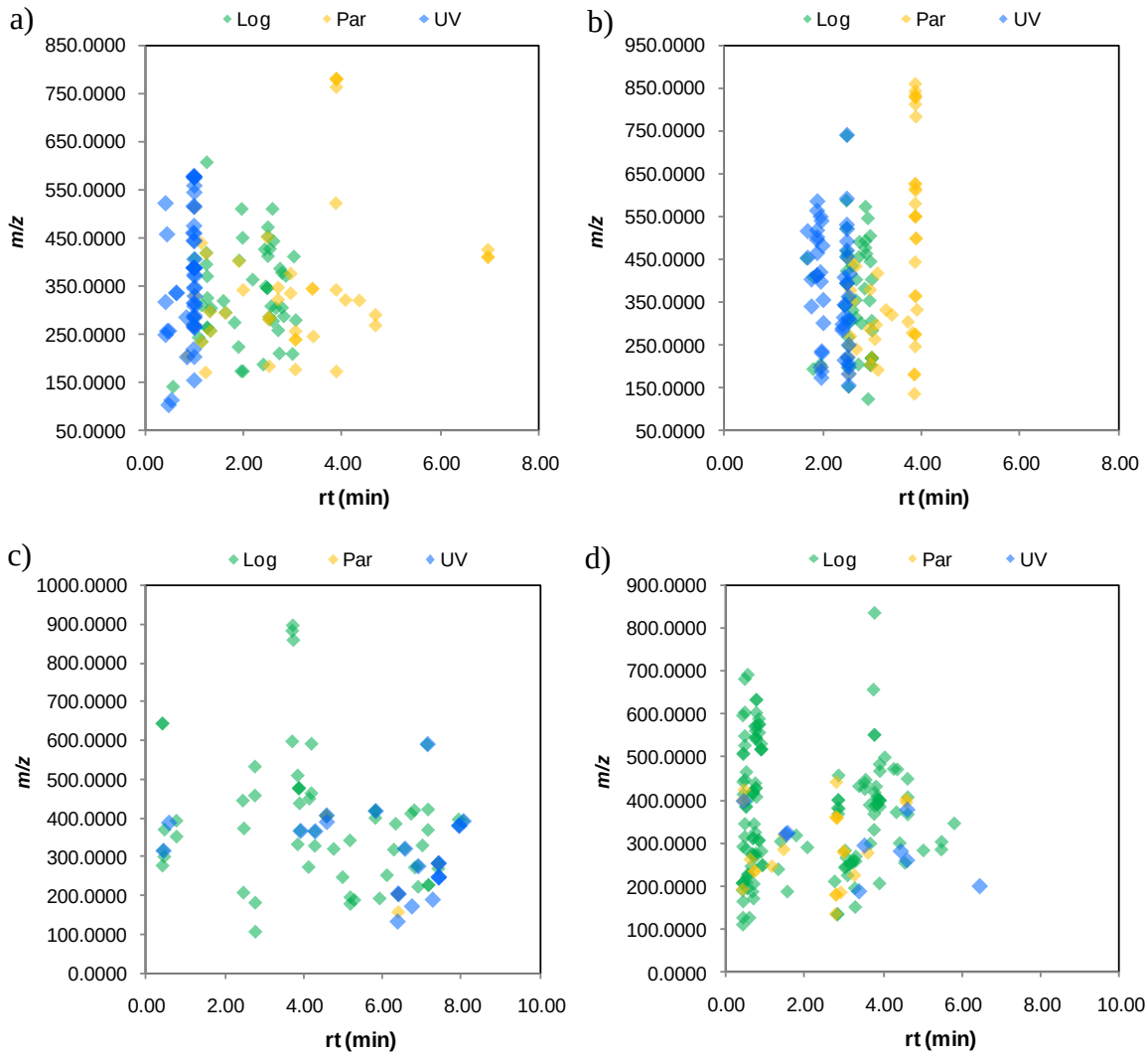


Figure 6.7 Plots representing the variables selected by scaling/transformation for datasets a) HSS ESI+, b) HSS ESI-, c) HILIC ESI+ and d) HILIC ESI-.

The selection/identification of MS features relevant in cancer vs. control discrimination was then further pursued by carefully inspecting the extracted ion chromatogram and mass spectrum of each feature. Background peaks, isotopes and adducts of assigned compounds were removed from the final list, so that the final numbers of selected features were: HSS ESI+ 118, HSS ESI- 113, HILIC ESI+ 66 and HILIC ESI- 147. The complete list of selected features is presented in Annex VII.

The tentative assignment of selected MS features to specific compounds was performed based on the comparison of experimental *m/z* and retention times with data for reference compounds compiled in the literature (Roux et al. 2012) and in a database available at Imperial College of London, where this analysis was performed. This strategy

allowed ten compounds to be tentatively identified (as listed in Table 6.3). One such example is phenylacetylglutamine, the corresponding extracted ion chromatogram and mass spectrum (obtained from HSS ESI+ detection) being illustrated in Figure 6.8.

Table 6.3 Assigned urinary UPLC-MS features found to be important in cancer vs. control discrimination. Features intensities are indicated as: very high (>10^5), high (>10^4, <10^5), low (>10^3, <10^4) and very low (<10^3).

Metabolite (adduct/fragment)	UPLC column & detection mode	m/z	RT (min)	feature intensity	PLS-DA scaling/ transfor-mation	% variation	*p*-value[d]	effect size±CI
Choline (M+)	HSS ESI+	104.10	0.50	very low	UV	26.5±6.0	$5.6x10^{-4}$	0.79±0.41
Creatinine (M+H)		114.07	0.55	high	UV	22.1±3.4	$8.2x10^{-8}$	1.18±0.43
Trigonelline (M+H)[c]		139.06	0.57	low	Log	-59.6±14.9	$7.6x10^{-8}$	-1.13±0.43
Pseudouridine (M+Na)[c]		269.07	0.99	low	Pareto	61.1±8.9	$3.4x10^{-7}$	1.08±0.42
Uric acid (M+H)		169.03	1.23	high	Pareto	48.4±8.8	$5.8x10^{-5}$	0.90±0.42
Pantothenic acid (M-H+HCOONa)[b]	HSS ESI-	286.09	2.94	low	Pareto	75.1±13.0	$1.8x10^{-4}$	0.85±0.41
Pyrocathecol sulfate isomer (M-H)[c]		189.99	3.11	low	Pareto	-47.7±11.4	$5.1x10^{-8}$	-1.09±0.42
Hippuric acid (M-H)-CO_2[a]		134.06	3.86	low	Pareto	-58.1±11.0	$1.7x10^{-11}$	-1.48±0.45
(M-H)[c]		179.06		low		-55.1±10.6	$1.8x10^{-10}$	-1.42±0.44
(M-H)[c]		180.06		low		-56.4±10.5	$7.3x10^{-11}$	-1.48±0.45
Phenylacetylglutamine (2M-2H+Na)[b]		549.19	3.88	low	Pareto	40.1±7.4	$1.6x10^{-5}$	0.93±0.42
(2M-2H+Na)[c]		550.20		high		41.3±7.7	$3.7x10^{-5}$	0.90±0.42
(2M-2H+Na)[c]		551.20		low		41.9±8.0	$3.7x10^{-5}$	0.89±0.42
Hippuric acid (M+H)-($C_2H_5NO_2$)[a]	HILIC ESI+	105.03	2.79	very low	Log	-66.5±17.3	$4.8x10^{-9}$	-1.13±0.43
(M+H)		180.07		low		-53.4±13.1	$4.9x10^{-10}$	-1.10±0.42
Hippuric acid (M-H)-CO_2[a]	HILIC ESI-	134.06	2.81	high	Pareto	-61.1±12.6	$1.5x10^{-11}$	-1.38±0.44
(M-H)[c]		179.06	2.82	very high		-53.7±10.4	$2.6\ x10^{-11}$	-1.40±0.44
(2M-H)[b]		357.11	2.84	high		-83.6±25.3	$3.2x10^{-6}$	-1.11±0.43
(2M-H)[c]		358.11	2.85	high		-82.4±24.3	$3.8x10^{-5}$	-1.12±0.43
(2M-2H+Na)[b]		379.09	2.87	high		-76.1±22.3	$8.7x10^{-7}$	-1.08±0.42
p-Hydroxyhippuric acid (M-H)-CO_2[a]		194.04	3.28	very high	Log	-47.2±20.7	$4.7x10^{-5}$	-0.60±0.40
(M-H)		150.05		low		-51.8±22.0	$2.4x10^{-9}$	-0.63±0.41
Phenylacetylglutamine (2M-2H+Na)[c]		551.21	3.77	low	Log	61.8±12.4	$7.4x10^{-9}$	0.78±0.41
(2M-2H+Na)[c]		552.21		low		93.0±14.0	$1.0x10^{-4}$	0.93±0.42

[a]Metabolite fragment. [b]Metabolite adduct. [c]Isotope of the molecular peak or adduct. [d]Wilcoxon rank sum test (p<0.01).

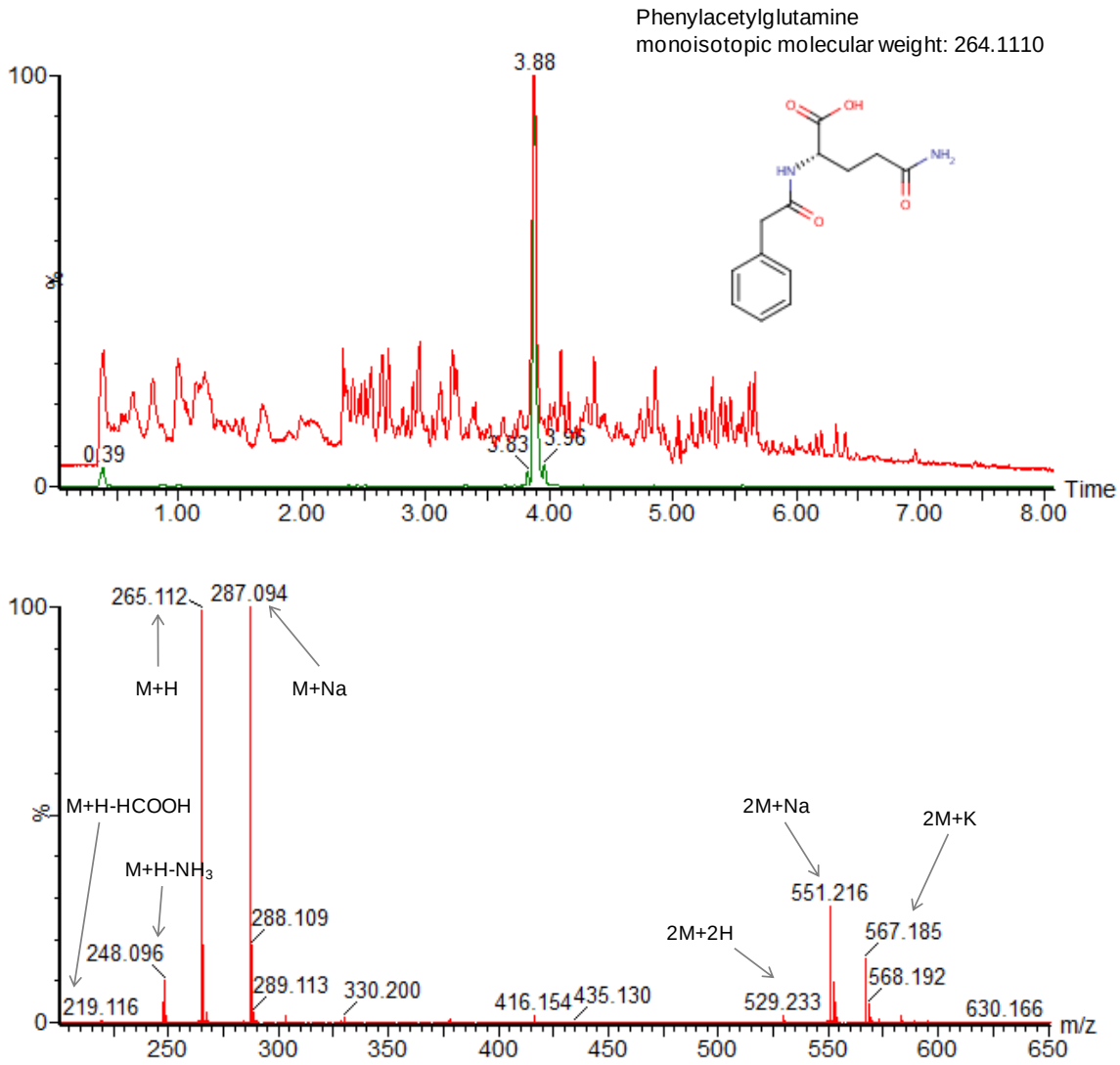


Figure 6.8 a) Total ion chromatogram (TIC) of representative urine and extracted ion chromatogram (EIC) of *m/z* 265.10. b) Mass spectrum of phenylacetylglutamine showing the main adducts formed.

In an attempt to improve the assignment of MS features, statistical correlation heterospectroscopy (SHY) of MS data with ^{1}H NMR spectra of the same samples was also explored. This approach allowed some assignments to be confirmed, namely those of hippurate, trigonelline and phenylacetylglutamine. The 2D correlation plot between UPLC-MS and ^{1}H NMR is illustrated in Figure 6.9a, for the HSS ESI- dataset, in which hippurate correlation in both domains is highlighted (Figure 6.9a left). This correlation was further confirmed by plotting hippurate MS intensities and NMR integral areas for the same samples, as illustrated in the scatter graph of Figure 6.9a right, for which a significant

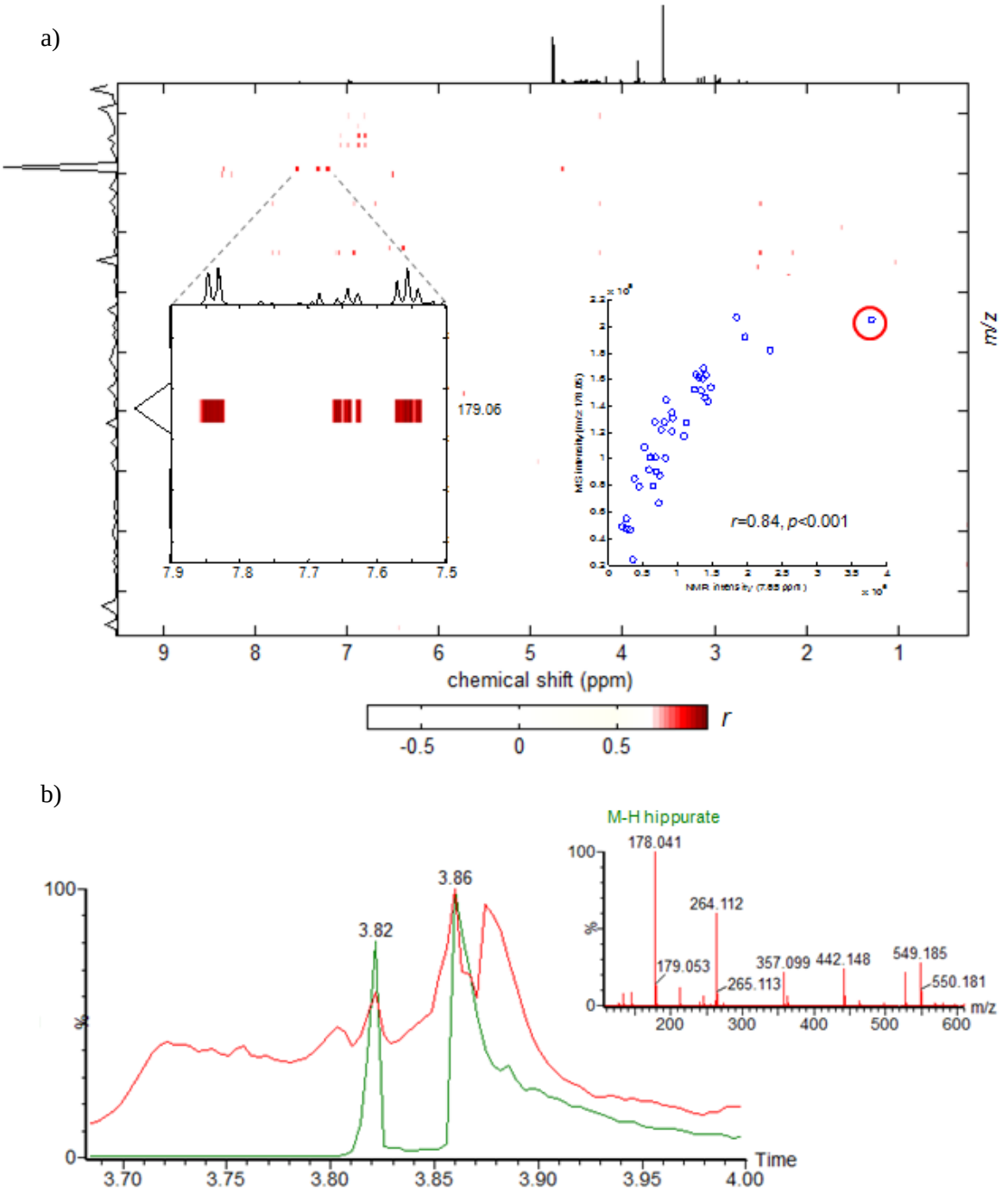


Figure 6.9 a) 2D statistical heterospectroscopy (SHY) map between UPLC-MS HSS ESI- (relevant features) and ^{1}H NMR (full resolution spectra) of the same urine samples. Correlation results are shown for hippurate, in which a sample deviating from linearity is highlighted. A threshold of r>0.7 and p<0.01 was applied. b) Expansion of the total ion chromatogram (TIC) of the same urine sample (red) showing the ion suppression effect observed for hippurate and corresponding extracted ion chromatogram (EIC) of *m/z* 179.04 (green, M-H hippurate).

Pearson correlation (r 0.84) was obtained. This graph also shows that one of the samples (highlighted with a red circle) deviated from the linear relationship observed for the other

samples. Ion suppression effects arising from the competition for ionization between the analyte of interest and other species may have contributed for this observation. Indeed, these effects, illustrated in Figure 6.9b, were also observed in several other cases (e.g., creatinine, citrate), which may help explaining the reduced number of significant correlations resulting from SHY analysis. Other than that, the different nature of the metabolites detected by the two methods (NMR and UPLC-MS) may also account for the observed inefficiency of SHY in further helping the assignment of MS features.

Amongst assigned metabolites, negative variations (decreased levels in patients' urine) were found for trigonelline, pyrocathecol sulphate, hippurate and *p*-hydroxyhippurate; whereas positive variations (increased levels in patients' urine) were found for choline, creatinine, pseudouridine, uric acid, pantothenic acid and phenylacetylglutamine. Interestingly, some of these variations, namely those of trigonelline, hippurate, creatinine and phenylacetylglutamine, were also detected by ^{1}H NMR profiling of urine, as presented in the previous chapter, while others were newly detected by UPLC-MS, thus showing the complementary nature of the two analytical platforms.

Compared to NMR, the more sensitive UPLC-MS methof highlighted a much higher number of discriminat features. However, their assignment to specific compounds was found to be a very difficult task, which could not be fully explored within the framework of this thesis. Future work should therefore entail the more thorough identification of MS features, namely with the help of MS/MS experiments to provide information on ion fragmentation patterns. Moreover, it would also be important to extend this analysis to an enlarged number of samples and to assess the influence of possible confounding factors.

7 FINAL CONCLUSIONS AND FUTURE PERSPECTIVES

In this work, the metabolic signature of lung cancer has been investigated through a metabolomics approach involving the analysis of preoperative tissues and biofluids (blood plasma and urine) from patients with primary lung tumours. An overview of the main variations entailing this signature, either considering the whole dataset, or each of the two main histological types included in this study, is presented in Figures 7.1 and 7.2, respectively.

The ^{1}H HRMAS NMR analysis of intact tissues (from 56 patients) enabled the composition of human lung tumours to be thoroughly characterized and offered a direct window into altered cellular metabolism. Over fifty metabolites have been identified, of which bonded amino acids (in peptides), polyamines and three other compounds were, to our knowledge, newly reported in this thesis, adding novel information about the metabolic composition of human lung tissue. Multivariate modelling of tissue spectra enabled lung tumours to be discriminated from adjacent, non-involved parenchyma with high accuracy (classification rate 97%) and predictive power (Q^2 0.83), as assessed by Monte Carlo cross validation (MCCV) of the classification model built. This discrimination was mainly due to significant differences in the levels of thirteen metabolites, namely glucose and acetate (decreased in tumours), together with lactate, alanine, glutamate, GSH, taurine, creatine, phospholipid-related compounds (PC, GPC and PE), uracil nucleotides and peptides (increased in tumours). While some of these differences corroborated typical features of cancer metabolism (e.g., upregulated glycolysis and glutaminolysis, altered lipid metabolism), other newly reported variations suggested less known pathways (e.g., antioxidant protection, protein degradation) to play important roles in lung cancer biology.

The tumour samples analysed differed in several parameters (as seen by microscopic observation of their mirror sections), namely the amount of tumour vs. stromal cells, the percentage of necrosis, the progression stage and the histological type of the tumour; hence, the possible influence of such parameters on the tumours' metabolic profile has been assessed. Interestingly, the variable proportion of tumour cells was found to be weakly correlated with the resulting spectral profiles, and the malignancy signature could be detected even in tissue samples with higher amounts of stroma. This result underlines the potential usefulness of real-time HRMAS NMR analysis of tissue fragments during

surgery to assist decisions on tumour resection margins and reduce the risk of local recurrence. In regard to the percentage of necrosis, it correlated with lipid abundance, which in turn had low importance in tumour vs. control discrimination. Concerning stage, tumour samples of stages I, II and III could not be distinguished, as they shared a similar discriminant signature in relation to control tissues, meaning that metabolic alterations were present even in small, locally confined tumours, showing little change as the tumour progressed. However, this statement should not be generalized as the large majority of tumours included in this study were from stages I and II, with only six cases of stage III and no stage IV samples. The tumour histological type, on the other hand, was found to be strongly reflected on the metabolic profiles measured. In particular, although having many variations in common, the two most prevalent subtypes – adenocarcinoma (AdC) and squamous cell carcinoma (SqCC) – showed distinct metabolic behaviour, especially when considering the inter-metabolite correlation patterns. Major alterations in AdC were related to phospholipid metabolism (as seen by elevated levels of PC, GPC and PE) and protein catabolism (increased peptide moieties), whereas SqCC had stronger glycolytic and glutaminolytic profiles (as shown by the negatively correlated variations in glucose and lactate and the positively correlated increases in glutamate and alanine). These results provided new, clear evidence of distinct metabolic signatures for AdC and SqCC, which could have important implications in the selective definition of new therapy targets or imaging tracers. Furthermore, the two subtypes could be discriminated with high accuracy, based on a set of six metabolites (mainly related to lipid metabolism and oxidative protection), thus showing that NMR metabolomics may indeed be valuable in tumour subtyping, which is a critical diagnostic requirement.

Based on the hypothesis that altered tumour metabolism may impact systemic metabolism and be reflected on the composition of biofluids, blood plasma and urine from over 100 patients (including those for which tissue samples were available) and near 100 controls were analysed by ^{1}H NMR spectroscopy. Multivariate analysis of plasma spectra allowed a robust classification model to be built, which discriminated lung cancer and healthy subjects with about 86% sensitivity and 88% specificity. The main discriminant features in the patients' plasma were found to be increased levels of lactate, pyruvate and acetoacetate, together with decreased levels of several amino acids (glutamine, histidine, valine, arginine/lysine and serine), methanol and two unassigned compounds. Moreover,

compared to controls, the plasma of patients was characterized by relatively higher levels of LDL+VLDL lipoproteins and *N*-acetylated glycoproteins, together with reduced levels of HDL lipoproteins and, possibly, albumin (this latter variation not being confirmed, as albumin signals were too broad to be unequivocally assigned and integrated). Besides confirming some of the suspected metabolic alterations in lung cancer, also reported by others and relating to glycolysis, amino acids and lipoproteins metabolism, these findings also provided new clues on the possible importance of fatty acid oxidation and protein glycosylation. Also notably, all of these changes were present from initial disease stages and some of them, like increased lactate and pyruvate or decreased glutamine, nicely matched the variations observed in tissues. Regarding the dependency of plasma alterations on tumour histological type, only subtle differences were noted: AdC additionally showed increased levels of β-hydroxybutyrate and a more relevant valine decrease (relatively to controls), while the plasma of SqCC patients showed a significant decrease in alanine (not observed in AdC). The distinction between AdC and SqCC through plasma NMR metabolomics was however not possible. Likewise, no clear differentiation was found between disease stages.

The classification model built with urine NMR spectra afforded even better discrimination between lung cancer patients and healthy controls, as shown by the high sensitivity and specificity levels obtained (97% and 96%, respectively). In this case, the list of metabolites contributing for this discrimination was more extensive and comprised β-hydroxybutyrate, α-hydroxyisobutyrate, allantoin, methylguanidine, *p*-cresol sulphate, six *N*-acetylated metabolites and 3 unknowns (increased in patients), together with trigonelline, tartrate, hippurate, 4-deoxythreonate and fatty acids (decreased in patients). Interestingly, these changes, detected from initial disease stages, corroborated the possible importance of biochemical events like fatty acid oxidation, oxidative stress and protein metabolism, also highlighted through tissue and/or plasma analysis (even if based on changes in different metabolites). Moreover, urinary profiling suggested a number of compounds typically associated with diet or gut microflora to be relevant in the cancer signature, thus reinforcing the increasingly recognised idea that gut microbiome-host interactions may play an important role in cancer pathogenesis and progression. Regarding patient discrimination by histological type or stage, urinary profiles seemed to comprise subtle differences that could be captured by multivariate analysis, resulting in MCCV-

validated models with modest predictive ability, although individual marker metabolites could not be identified.

One of the limitations of this study, in what concerns the analysis of biofluids, was the imperfect matching between cancer and control groups in terms of demographic and environmental factors. For instance, the cancer group comprised more males and older subjects than the control group. To deal with this problem, the possible influence of a few unmatched parameters, namely gender, age and smoking habits, on cancer vs. control discrimination was assessed. As previously reported in the literature, the levels of some metabolites in both biofluids were found to be gender-dependent (as assessed within controls). Hence, by confronting the list of gender-related variables with that of putative cancer markers, possibly biased variations (i.e., variations showing the same direction in male and cancer groups) have been highlighted and their importance in cancer vs. control discrimination re-assessed by considering separate models for males and females. In the case of plasma, all previously identified discriminant features remained important regardless of gender, whereas, in urine, creatine and creatinine were found to be gender-biased, as they were no longer significantly different between controls and patients, when considering each gender separately. A similar strategy has been devised to evaluate the possibly confounding effect of age mismatch between control and cancer groups. After identifying age-related variables in the control group, for both biofluids, the metabolites found to be simultaneously increased or decreased in both older subjects and patients were highlighted as possibly biased and their importance as putative cancer markers was reassessed in a subset of samples having better age-matching between control and cancer groups. The resulting plasma-based classification model decreased its MCCV quality parameters; therefore, although all previously identified discriminant metabolites in plasma held their importance in the age-matched model, age-bias could not be ruled out as an important influencing factor. On the other hand, in the case of urine, the age-matched model maintained high accuracy and predictive power, thus allowing age to be excluded as a strong confounding factor. Still, a few metabolites, namely citrate, histidine and phenylacetylglutamine, were found to be age-biased (as they were not important for cancer vs. control discrimination in the new age-matched model). In regard to the influence of smoking, no relation was found between smoke exposure and the plasma profile of control subjects, whereas some urinary metabolites (mostly unknowns) seemed to be associated

with smoking habits. Nevertheless, discrimination between controls and cancer patients was equally good when the control group comprised both smokers and never smokers or each of these subsets separately, thus showing the negligible influence of smoking habits on cancer vs. control discrimination.

Following the analysis of possible confounders and the exclusion of possibly biased variables (in the case of urine), the classification ability of multivariate models was further evaluated by preliminary external validation. For this purpose, the sample cohort was divided into a validation set used for model building and a prediction set comprising 40 samples collected more recently in time. In the case of plasma, the best external prediction was achieved when using the integrals of twelve metabolites, previously identified as cancer-related markers (rather than using the whole spectral profile). Only three out of twenty controls and three out of twenty patients were misclassified (85% sensitivity and specificity), which is a promising result, especially considering that currently explored blood markers, like the carcinoembryonic antigen (CEA) and the cytokeratin 19 fragment CYFRA 21-1, have lower performance. The results of external validation for urinary markers were even more exciting as all patients were correctly classified and only two controls were misclassified as cancer cases. Although being exploratory and requiring further confirmation in an enlarged sample set, with improved control over possible confounders, the novelty and possible clinical impact of these findings, together with the easiness and non-invasiveness of urine collection, make it extremely encouraging to pursue further work on the validation of a urinary signature for lung cancer. For instance, an envisaged useful application would be to use the urinary metabolic profile (eventually in tandem with the plasma profile) as a pre-screening tool for selecting subjects who should undergo more specific or advanced radiological testing, namely within population groups at increased risk (e.g., chronic smokers).

In face of the promising results obtained by NMR profiling of urine (described above), the use of an alternative, complementary profiling technique (UPLC-MS) has also been explored for a subset of samples. The classification models built (for data obtained with different chromatographic columns, ionization modes and processing methods) showed sensitivities higher than 87% and specificities above 89% for cancer vs. control discrimination, which corroborates the value of urine as a rich source of disease markers. Over 600 MS features were selected as being important for class discrimination, although

only 10 of those features could be assigned to specific metabolites, thus underlying the importance of carrying out further studies (namely MS/MS analysis) to improve the structural information retrieved. Identified metabolites were choline, creatinine, pseudouridine, uric and pantothenic acids and phenylacetylglutamine (increased in patients), together with trigonelline, pyrocathecol sulphate, hippurate and *p*-hydroxyhippurate (decreased in patients). Compared to ^{1}H NMR profiling of the same samples, common variations were observed for trigonelline, hippurate, phenylacetylglutamine and creatinine, although the last two metabolites were found to be age and gender-related, respectively.

Overall, the metabolomics approach used in this thesis enabled a number of significant metabolite variations to be highlighted as part of a putative lung cancer metabolic signature. To our knowledge, this is the first time that the information derived from three biological matrices (tissues, blood plasma and urine) is brought together for defining a global signature for lung cancer; the complementary nature of the metabolic information revealed in this way, even if not completely understood at the moment, underlines the importance of such an integrative approach. In regard to the validation of this signature, which is crucial to drive the results of this work into clinical applications that can improve patients care, it would be important to pursue future work along the following avenues: i) enlargement of the sample set in order to verify the robustness of findings; ii) modelling of possible confounders not considered in this work, such as body mass, diet, comorbidities (e.g., diabetes, cerebrovascular and peripheral vascular diseases); iii) more thorough investigation of the dependence on the histological subtype, with a view to develop new tools for the differential diagnosis of lung tumours; iv) assessment of the specificity of the metabolic alterations found to be associated with lung cancer, as compared to other lung diseases (e.g., chronic obstructive pulmonary disease, asthma, fibrosis, pneumonia, tuberculosis); v) biological validation of the metabolic alterations observed, for instance by measuring the expression/activity of specific enzymes and by performing stable isotope-resolved metabolomics (SIRM) to follow the fate of selected substrates; vi) correlation of metabolomics results with proteomics and genomics data, with a view to achieve a more comprehensive understanding of lung cancer biology.

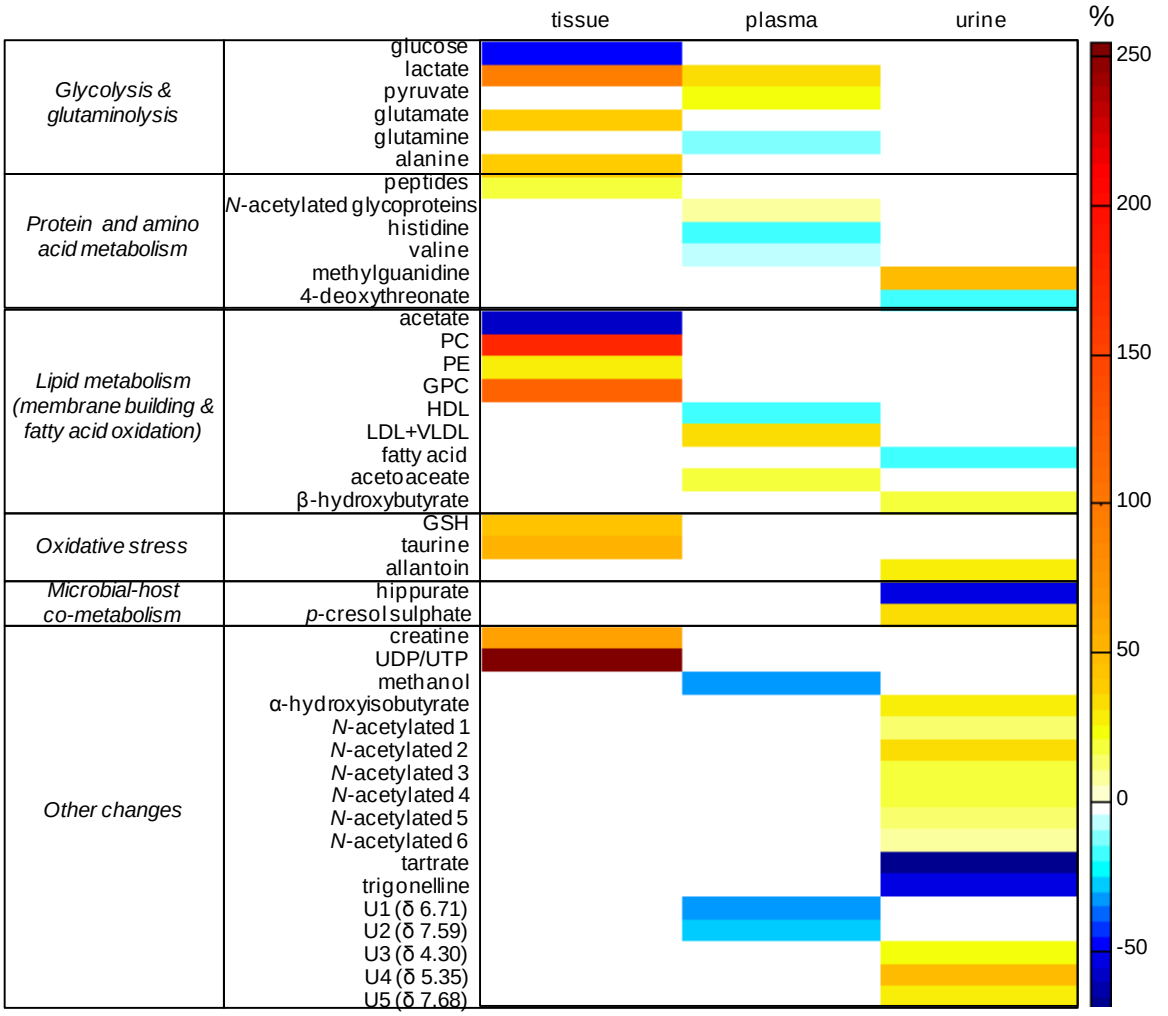


Figure 7.1 Heatmap of main metabolite variations found to differentiate lung tumours from control tissues or lung cancer patients from healthy controls (considering the whole dataset). The colour scale reflects the direction and magnitude of these variations (%) and the metabolites are grouped according to hypothesized altered metabolic pathways. Only the metabolites highlighted through NMR profiling, confirmed to show statistically significant differences in their levels and to be free of age- or gender-bias are included. GSH: reduced glutathione; GPC: glycerophosphocholine; HDL: high-density lipoproteins; LDL: low-density lipoprotein; PC, phosphocholine; PE: phosphoethanolamine; UDP/UTP: uridine di/triphosphate; VLDL: very low-density lipoprotein.

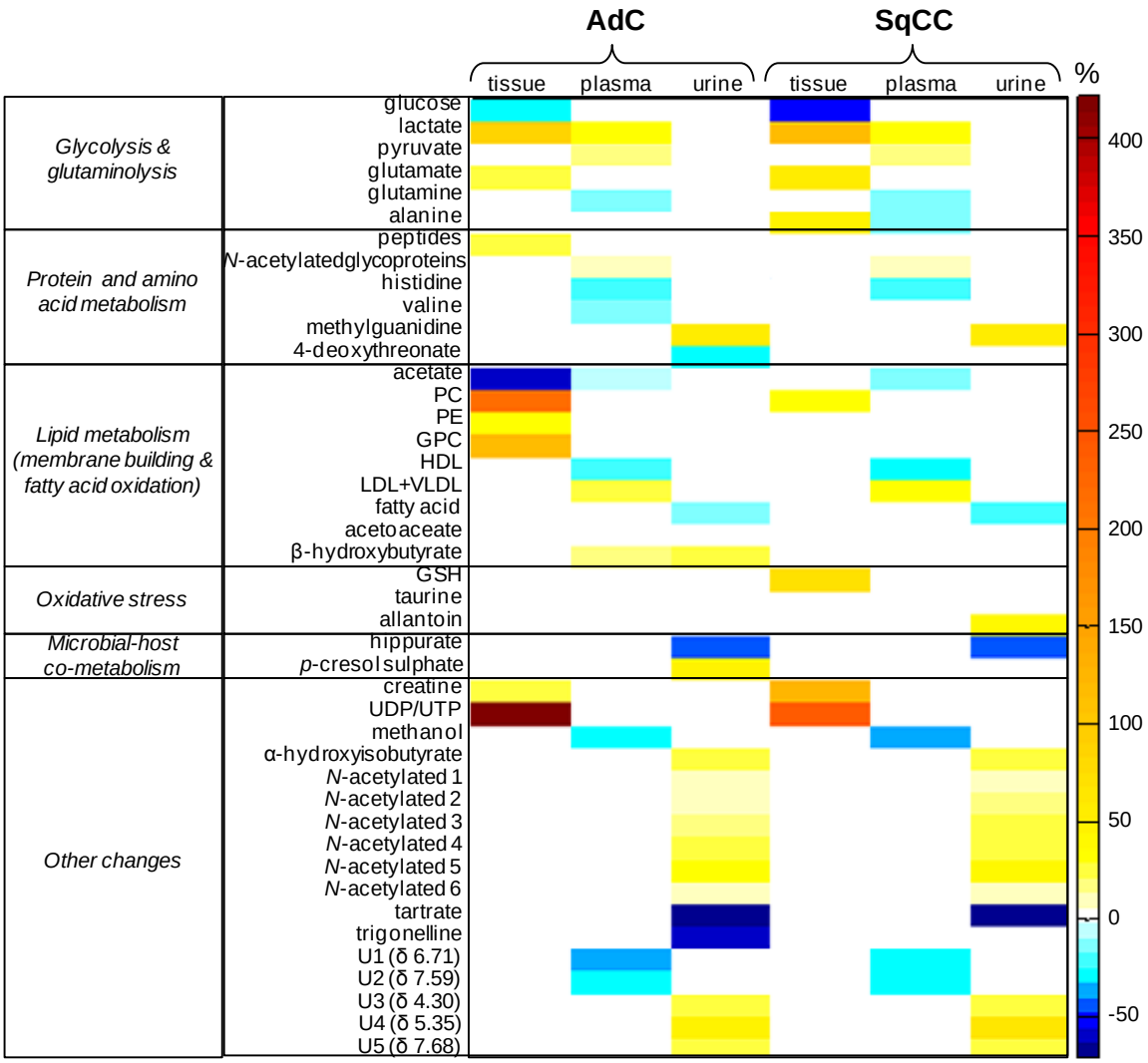


Figure 7.2 Heatmap of main metabolite variations found to differentiate lung tumours from control tissues or lung cancer patients from healthy controls, considering each of the two main histological types (AdC – adenocarcinoma and SqCC – squamous cell carcinoma). The colour scale reflects the direction and magnitude of these variations (%) and the metabolites are grouped according to hypothesized altered metabolic pathways. Only the metabolites highlighted through NMR profiling, confirmed to show statistically significant differences in their levels and to be free of age- or gender-bias are included. GSH: reduced glutathione; GPC: glycerophosphocholine; HDL: high-density lipoproteins; LDL: low-density lipoprotein; PC, phosphocholine; PE: phosphoethanolamine; UDP/UTP: uridine di/triphosphate; VLDL: very low-density lipoprotein.

BIBLIOGRAPHY

Aberle, D.R., Abtin, F. and Brown, K., 2013. Computed tomography screening for lung cancer: has it finally arrived? Implications of the national lung screening trial. *Journal of Clinical Oncology*, 31(8), pp.1002–8.

Aberle, D.R., Adams, A.M., Berg, C.D., Black, W.C., Clapp, J.D., Fagerstrom, R.M., Gareen, I.F., Gatsonis, C., Marcus, P.M. and Sicks, J.D., 2011. Reduced lung-cancer mortality with low-dose computed tomographic screening. *The New England Journal of Medicine*, 365(5), pp.395–409.

Aberle, D.R., Berg, C.D., Black, W.C., Church, T.R., Fagerstrom, R.M., Galen, B., Gareen, I.F., Gatsonis, C., Goldin, J., Gohagan, J.K., Hillman, B., Jaffe, C., Kramer, B.S., Lynch, D., Marcus, P.M., Schnall, M., Sullivan, D.D.C. and Zylak, C.J., 2011. The National Lung Screening Trial: overview and study design. *Radiology*, 258(1), pp.243–53.

Addis, B.J., Dewar, A. and Thurlow, N.P., 1988. Giant cell carcinoma of the lung-immunohistochemical and ultrastructural evidence of differentiation. *The Journal of Pathology*, 155(3), pp.231–40.

Ahluwalia, G.S., Grem, J.L., Hao, Z. and Cooney, D.A., 1990. Metabolism and action of amino acid analog anti-cancer agents. *Pharmacology & Therapeutics*, 46(2), pp.243–71.

Ala-Korpela, M., 1995. 1H NMR spectroscopy of human blood plasma. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 27(5), pp.475–54.

Alberg, A.J., Wallace, K., Silvestri, G.A. and Brock, M. V, 2013. Invited commentary: the etiology of lung cancer in men compared with women. *American Journal of Epidemiology*, 177(7), pp.613–6.

Ambrosini, V., Nicolini, S., Caroli, P., Nanni, C., Massaro, A., Marzola, M.C., Rubello, D. and Fanti, S., 2012. PET/CT imaging in different types of lung cancer: an overview. *European Journal of Radiology*, 81(5), pp.988–1001.

An, Z.L., Chen, Y.H., Zhang, R.P., Song, Y.M., Sun, J.H., He, J.M., Bai, J.F., Dong, L.J., Zhan, Q.M. and Abliz, Z., 2010. Integrated ionization approach for RRLC-MS/MS-based metabonomics: finding potential biomarkers for lung cancer. *Journal of Proteome Research*, 9(8), pp.4071–81.

Andersen, C.M. and Bro, R., 2010. Variable selection in regression -a tutorial. *Journal of Chemometrics*, 24(11-12), pp.728–37.

Andrew, E.R. and Newing, R.A., 1958. The narrowing of nuclear magnetic resonance spectra by molecular rotation in solids. *Proceedings of the Physical Society*, 72(6), pp.959–72.

Di Anibal, C. V, Callao, M.P. and Ruisánchez, I., 2011. 1H NMR variable selection approaches for classification. A case study: the determination of adulterated foodstuffs. *Talanta*, 86, pp.316–23.

Appiah-Amponsah, E., Shanaiah, N., Nagana Gowda, G.A., Owusu-Sarfo, K., Ye, T. and Raftery, D., 2009. Identification of 4-deoxythreonic acid present in human urine using HPLC and NMR techniques. *Journal of Pharmaceutical and Biomedical Analysis*, 50(5), pp.878–85.

Ardenkjaer-Larsen, J.H., Fridlund, B., Gram, A., Hansson, G., Hansson, L., Lerche, M.H., Servin, R., Thaning, M. and Golman, K., 2003. Increase in signal-to-noise ratio of > 10,000 times in liquid-state NMR. *Proceedings of the National Academy of Sciences of the United States of America*, 100(18), pp.10158–63.

Asamura, H., Kameya, T., Matsuno, Y., Noguchi, M., Tada, H., Ishikawa, Y., Yokose, T., Jiang, S.-X., Inoue, T., Nakagawa, K., Tajima, K. and Nagai, K., 2006. Neuroendocrine neoplasms of the lung: a prognostic spectrum. *Journal of Clinical Oncology*, 24(1), pp.70–6.

Attanoos, R.L., Papagiannis, A., Suttinont, P., Goddard, H., Papotti, M. and Gibbs, A.R., 1998. Pulmonary giant cell carcinoma: pathological entity or morphological phenotype? *Histopathology*, 32(3), pp.225–31.

Bach, P.B., 2003. Screening for lung cancer. *CHEST Journal*, 123(Suppl 1), pp.S72-82.

Bach, P.B., Jett, J.R., Pastorino, U., Tockman, M.S., Swensen, S.J. and Begg, C.B., 2007. Computed tomography screening and lung cancer outcomes. *Journal of the American Medical Association*, 297(9), pp.953–61.

Banerjee, S. and Shyamalava Mazumdar, 2012. Electrospray ionization mass spectrometry: a technique to access the information beyond the molecular weight of the analyte. *International Journal of Analytical Chemistry*, 2012, article ID 282574, 40 pages.

Bartella, L. and Huang, W., 2007. Proton (1H) MR spectroscopy of the breast. *Radiographics*, 27 (Suppl 1), pp.S241–52.

Bax, A. and Davis, D.G., 1985. MLEV-17-based two-dimensional homonuclear magnetization transfer spectroscopy. *Journal of Magnetic Resonance*, 65(2), pp.355–60.

Beckonert, O., Coen, M., Keun, H.C., Wang, Y., Ebbels, T.M.D., Holmes, E., Lindon, J.C. and Nicholson, J.K., 2010. High-resolution magic-angle-spinning NMR spectroscopy for metabolic profiling of intact tissues. *Nature Protocols*, 5(6), pp.1019–32.

Beckonert, O., Keun, H.C., Ebbels, T.M.D., Bundy, J., Holmes, E., Lindon, J.C. and Nicholson, J.K., 2007. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature Protocols*, 2(11), pp.2692–703.

Bell, J.D., Brown, J.C.C., Nicholson, J.K. and Sadler, P.J., 1987. Assignment of resonances for acute-phase glycoproteins in high resolution proton NMR spectra of human blood plasma. *FEBS Letters*, 215(2), pp.311–15.

Benahmed, M.A., Elbayed, K.,Daubeuf, F., Santelmo, N., Frossard, N., Namer, I.J., 2014. NMR HRMAS spectroscopy of lung biopsy samples: Comparison study between human, pig, rat, and mouse metabolomics. *Magnetic Resonance in Medicine*, 71(1), pp.35-43.

Bensaad, K. and Vousden, K.H., 2007. p53: new roles in metabolism. *TRENDS in Cell Biology*, 17(6), pp.286–91.

Berben, L., Sereika, S.M. and Engberg, S., 2012. Effect size estimation: methods and examples. *International Journal of Nursing Studies*, 49(8), pp.1039–47.

Bharti, S.K. and Roy, R., 2012. Quantitative 1H NMR spectroscopy. *Trends in Analytical Chemistry*, 35, pp.5–26.

Bijlsma, S., Bobeldijk, I., Verheij, E.R., Ramaker, R., Kochhar, S., Macdonald, I.A., van Ommen, B. and Smilde, A.K., 2005. Large-scale human metabolomics studies: a strategy for data (pre-)processing and validation. *Analytical Chemistry*, 78(2), pp.567–74.

Blair, S.L., Heerdt, P., Sachar, S., Abolhoda, A., Hochwald, S., Cheng, H. and Burt, M., 1997. Glutathione metabolism in patients with non-small cell lung cancers. *Cancer Research*, 57(1), pp.152–5.

Bleeker, S.. E., Moll, H.. A., Steyerberg, E.. W., Donders, A.R.. R.T., Derksen-Lubsen, G., Grobbee, D.. E. and Moons, K.G.. G.M., 2003. External validation is necessary in prediction research: *Journal of Clinical Epidemiology*, 56(9), pp.826–32.

Bodenhausen, G. and Ruben, D.J., 1980. Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. *Chemical Physics Letters*, 69(1), pp.185–9.

Bollard, M.E., Garrod, S., Holmes, E., Lindon, J.C., Humpfer, E., Spraul, M. and Nicholson, J.K., 2000. High-resolution (1)H and (1)H-(13)C magic angle spinning NMR spectroscopy of rat liver. *Magnetic Resonance in Medicine*, 44(2), pp.201–7.

Boros, L.G., Lee, P.W.N., Brandes, J.L., Cascante, M., Muscarella, P., Schirmer, W.J., Melvin, W.S. and Ellison, E.C., 1998. Nonoxidative pentose phosphate pathways and their direct role in ribose synthesis in tumors: is cancer a disease of cellular glucose metabolism? *Medical Hypotheses*, 50(1), pp.55–9.

Bouatra, S., Aziat, F., Mandal, R., Guo, A.C., Wilson, M.R., Knox, C., Bjorndahl, T.C., Krishnamurthy, R., Saleem, F., Liu, P., Dame, Z.T., Poelzer, J., Huynh, J., Yallou, F.S., Psychogios, N., Dong, E., Bogumil, R., Roehring, C. and Wishart, D.S., 2013. The human urine metabolome. *PLoS ONE*, 8(9), article ID e73076, 28 pages.

Boyle, P. and Levin, B., 2009. *World Cancer Report 2008* 1st Edition, World Health Organization.

Bray, F., Ren, J.-S., Masuyer, E. and Ferlay, J., 2013. Global estimates of cancer prevalence for 27 sites in the adult population in 2008. *International Journal of Cancer*, 132(5), pp.1133–45.

Broadhurst, D.I. and Kell, D.B., 2006. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics*, 2(4), pp.171–96.

Brothers, J.F., Hijazi, K., Mascaux, C., El-Zein, R.A., Spitz, M.R. and Spira, A., 2013. Bridging the clinical gaps: genetic, epigenetic and transcriptomic biomarkers for the early detection of lung cancer in the post-National Lung Screening Trial era. *BMC Medicine*, 11, article ID 168, 15 pages.

Brown, R.S., Leung, J.Y., Kison, P. V, Zasadny, K.R., Flint, A. and Wahl, R.L., 1999. Glucose transporters and FDG uptake in untreated primary human non-small cell lung cancer. *Journal of Nuclear Medicine*, 40(4), pp.556–65.

Brusselmans, K., De Schrijver, E., Verhoeven, G. and Swinnen, J. V, 2005. RNA interference-mediated silencing of the Acetyl-CoA-Carboxylase-alpha gene induces

growth inhibition and apoptosis of prostate cancer cells. *Cancer Research*, 65(15), pp.6719–25.

Bullinger, D., Fux, R., Nicholson, G., Plontke, S., Belka, C., Laufer, S., Gleiter, C.H. and Kammerer, B., 2008. Identification of urinary modified nucleosides and ribosylated metabolites in humans via combined ESI-FTICR MS and ESI-IT MS analysis. *Journal of the American Society for Mass Spectrometry*, 19(10), pp.1500–13.

Bultitude, F.W. and Newham, S.J., 1975. Identification of some abnormal metabolites in plasma from uremic subjects. *Clinical Chemistry*, 21(9), pp.1329–34.

Burns, T.R., Underwood, R.D., Greenberg, S.D., Teasdale, T.A. and Cartwright, J., 1989. Cytomorphometry of large cell carcinoma of the lung. *Analytical and Quantitative Cytology and Histology*, 11(1), pp.48–52.

Cai, X., Dong, J., Zou, L., Xue, X., Zhang, X. and Liang, X., 2011. Metabonomic study of lung cancer and the effects of radiotherapy on lung cancer patients: analysis of highly polar metabolites by ultraperformance HILIC coupled with Q-TOF MS. *Chromatographia*, 74(5-6), pp.391–8.

Carracedo, A., Cantley, L.C. and Pandolfi, P.P., 2013. Cancer metabolism: fatty acid oxidation in the limelight. *Nature Reviews Cancer*, 13(4), pp.227–32.

Carrola, J., Rocha, C.M., Barros, A.S., Gil, A.M., Goodfellow, B.J., Carreira, I.M., Bernardo, J., Gomes, A., Sousa, V., Carvalho, L. and Duarte, I.F., 2011. Metabolic signatures of lung cancer in biofluids: NMR-based metabonomics of urine. *Journal of Proteome Research*, 10(1), pp.221–30.

Cascante, M., Centelles, J.J., Veech, R.L., Lee, W.N.P. and Boros, L.G., 2000. Role of thiamin (vitamin B-1) and transketolase in tumor cell proliferation. *Nutrition and Cancer*, 36(2), pp.150–4.

Cavill, R., Keun, H.C., Holmes, E., Lindon, J.C., Nicholson, J.K. and Ebbels, T.M.D., 2009. Genetic algorithms for simultaneous variable and sample selection in metabonomics. *Bioinformatics*, 25(1), pp.112–8.

Chan, E.C.Y., Koh, P.K., Mal, M., Cheah, P.Y., Eu, K.W., Backshall, A., Cavill, R., Nicholson, J.K. and Keun, H.C., 2009. Metabolic profiling of human colorectal cancer using high-resolution magic angle spinning nuclear magnetic resonance (HR-MAS NMR) spectroscopy and gas chromatography mass spectrometry (GC/MS). *Journal Proteome Research*, 8(1), pp.352–61.

Chaneton, B., Hillmann, P., Zheng, L., Martin, A.C.L., Maddocks, O.D.K., Chokkathukalam, A., Coyle, J.E., Jankevics, A., Holding, F.P., Vousden, K.H., Frezza, C., O'Reilly, M. and Gottlieb, E., 2012. Serine is a natural ligand and allosteric activator of pyruvate kinase M2. *Nature*, 491(7424), pp.458–62.

Chaudhri, V.K., Salzler, G.G., Dick, S.A., Buckman, M.S., Sordella, R., Karoly, E.D., Mohney, R., Stiles, B.M., Elemento, O., Altorki, N.K. and McGraw, T.E., 2013. Metabolic alterations in lung cancer-associated fibroblasts correlated with increased glycolytic metabolism of the tumor. *Molecular Cancer Research*, 11(6), pp.579–92.

Chejfec, G., Candel, A., Jansson, D.S., Warren, W.H., Koukoulis, G.K., Gould, J.E., Manderino, G.L., Gooch, G.T. and Gould, V.E., 1991. Immunohistochemical features of giant cell carcinoma of the lung: patterns of expression of cytokeratins, vimentin, and the mucinous glycoprotein recognized by monoclonal antibody A-80. *Ultrastructural Pathology*, 15(2), pp.131–8.

Chen, W., Zu, Y., Huang, Q., Chen, F., Wang, G., Lan, W., Bai, C., Lu, S., Yue, Y. and Deng, F., 2011. Study on metabonomic characteristics of human lung cancer using high resolution magic-angle spinning 1H NMR spectroscopy and multivariate data analysis. *Magnetic Resonance in Medicine*, 66(6), pp.1531–40.

Chen, Y.-J., Wang, X.-H., Huang, Z.-Z., Lin, L., Gao, Y., Zhu, E.-Y., Xing, J.-C., Zheng, J.-X. and Hang, W., 2012. A study of human bladder cancer by serum and urine metabonomics. *Chinese Journal of Analytical Chemistry*, 40(9), pp.1322–8.

Cheng, T., Sudderth, J., Yang, C., Mullen, A.R., Jin, E.S., Matés, J.M. and DeBerardinis, R.J., 2011. Pyruvate carboxylase is required for glutamine-independent growth of tumor cells. *Proceedings of the National Academy of Sciences of the United States of America*, 108(21), pp.8674–9.

Cheng, Y., Xie, G., Chen, T., Qiu, Y., Zou, X., Zheng, M., Tan, B., Feng, B., Dong, T., He, P., Zhao, L., Zhao, A., Xu, L.X., Zhang, Y. and Jia, W., 2012. Distinct urinary metabolic profile of human colorectal cancer. *Journal of Proteome Research*, 11(2), pp.1354–63.

Christiansen, M.N., Chik, J., Lee, L., Anugraham, M., Abrahams, J.L. and Packer, N.H., 2014. Cell surface protein glycosylation in cancer. *Proteomics*, 14(4-5), pp.525–46.

Ciebiada, M., Gorski, P. and Antczak, A., 2012. Eicosanoids in exhaled breath condensate and bronchoalveolar lavage fluid of patients with primary lung cancer. *Disease Markers*, 32(5), pp.329–35.

Claridge, T., 2000. *High-Resolution NMR Techniques in Organic Chemistry, Volume 19 (Tetrahedron Organic Chemistry)* 1st Edition, Oxford: Pergamon.

Clem, B., Telang, S., Clem, A., Yalcin, A., Meier, J., Simmons, A., Rasku, M.A., Arumugam, S., Dean, W.L., Eaton, J., Lane, A., Trent, J.O. and Chesney, J., 2008. Small-molecule inhibition of 6-phosphofructo-2-kinase activity suppresses glycolytic flux and tumor growth. *Molecular Cancer Therapeutics*, 7(1), pp.110–20.

Cloarec, O., Dumas, M.-E., Craig, A., Barton, R.H., Trygg, J., Hudson, J., Blancher, C., Gauguier, D., Lindon, J.C., Holmes, E. and Nicholson, J., 2005. Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic 1H NMR data sets. *Analytical Chemistry*, 77(5), pp.1282–9.

Cook, J.A., Pass, H.I., Iype, S.N., Friedman, N., DeGraff, W., Russo, A. and Mitchell, J.B., 1991. Cellular glutathione and thiol measurements from surgically resected human lung tumor and normal lung tissue. *Cancer Research*, 51(16), pp.4287–94.

Cooper, W.A., O'toole, S., Boyer, M., Horvath, L., Mahar, A., 2011. What's new in non-small cell lung cancer for pathologists: the importance of accurate subtyping, EGFR mutations and ALK rearrangements. *Pathology*, 43(2), pp.103-15.

Costello, L.C. and Franklin, R.B., 2005. "Why do tumour cells glycolyse?": From glycolysis through citrate to lipogenesis. *Molecular and Cellular Biochemistry*, 208(1-2), pp.1–8.

Coy, J.F., Dressler, D., Wilde, J. and Schubert, P., 2005. Mutations in the transketolase-like gene TKTL1: Clinical implications for neurodegenerative diseases, diabetes and cancer. *Clinical Laboratory*, 51(5-6), pp.257–73.

Craig, A., Cloarec, O., Holmes, E., Nicholson, J.K. and Lindon, J.C., 2006. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Analytical Chemistry*, 78(7), pp.2262–7.

Crockford, D.J., Holmes, E., Lindon, J.C., Plumb, R.S., Zirah, S., Bruce, S.J., Rainville, P., Stumpf, C.L. and Nicholson, J.K., 2006. Statistical heterospectroscopy, an approach to the integrated analysis of NMR and UPLC-MS data sets: application in metabonomic toxicology studies. *Analytical Chemistry*, 78(2), pp.363–71.

Crockford, D.J., Maher, A.D., Ahmadi, K.R., Barrett, A., Plumb, R.S., Wilson, I.D. and Nicholson, J.K., 2008. 1H NMR and UPLC-MSE statistical heterospectroscopy: characterization of drug metabolites (xenometabolome) in epidemiological studies. *Analytical Chemistry*, 80(18), pp.6835–44.

Dang, C. V and Semenza, G.L., 1999. Oncogenic alterations of metabolism. *Trends in Biochemical Sciences*, 24(2), pp.68–72.

Davis, V.W., Bathe, O.F., Schiller, D.E., Slupsky, C.M. and Sawyer, M.B., 2011. Metabolomics and surgical oncology: Potential role for small molecule biomarkers. *Journal of Surgical Oncology*, 103(5), pp.451–9.

Daye, D. and Wellen, K.E., 2012. Metabolic reprogramming in cancer: unraveling the role of glutamine in tumorigenesis. *Seminars in Cell & Developmental Biology*, 23(4), pp.362–9.

Daykin, C.A., Corcoran, O., Hansen, S.H., Bjørnsdottir, I., Cornett, C., Connor, S.C., Lindon, J.C. and Nicholson, J.K., 2001. Application of directly coupled HPLC NMR to separation and characterization of lipoproteins from human serum. *Analytical Chemistry*, 73(6), pp.1084–90.

DeBerardinis, R.J., Mancuso, A., Daikhin, E., Nissim, I., Yudkoff, M., Wehrli, S. and Thompson, C.B., 2007. Beyond aerobic glycolysis: transformed cells can engage in glutamine metabolism that exceeds the requirement for protein and nucleotide synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 104(49), pp.19345–50.

Detterbeck, F.C., Lewis, S.Z., Diekemper, R., Addrizzo-Harris, D. and Alberts, W.M., 2013. Executive summary: diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest*, 143(Suppl 5), pp.S7–37.

Diaz, S.O., Barros, A.S., Goodfellow, B.J., Duarte, I.F., Galhano, E., Pita, C., Almeida, M. do C., Carreira, I.M. and Gil, A.M., 2013. Second trimester maternal urine for the diagnosis of trisomy 21 and prediction of poor pregnancy outcomes. *Journal of Proteome Research*, 12(6), pp.2946–57.

Dieterle, F., Ross, A., Schlotterbeck, G. and Senn, H., 2006. Probabilistic quotient normalization as robust method to account for dilution of complex biological

mixtures. Application in 1H NMR metabonomics. *Analytical Chemistry*, 78(13), pp.4281–90.

Dong, J., Cai, X.M., Zhao, L.L., Xue, X.Y., Zou, L.J., Zhang, X.L. and Liang, X.M., 2010. Lysophosphatidylcholine profiling of plasma: discrimination of isomers and discovery of lung cancer biomarkers. *Metabolomics*, 6(4), pp.478–88.

Dong, J., Cheng, K.-K., Xu, J., Chen, Z. and Griffin, J.L., 2011. Group aggregating normalization method for the preprocessing of NMR-based metabolomic data. *Chemometrics and Intelligent Laboratory Systems*, 108(2), pp.123–32.

Dowling, C., Bollen, A.W., Noworolski, S.M., McDermott, M.W., Barbaro, N.M., Day, M.R., Henry, R.G., Chang, S.M., Dillon, W.P., Nelson, S.J. and Vigneron, D.B., 2001. Preoperative proton MR spectroscopic imaging of brain tumors: correlation with histopathologic analysis of resection specimens. *American Journal of Neuroradiology*, 22(4), pp.604–12.

Drilon, A., Rekhtman, N., Ladanyi, M. and Paik, P., 2012. Squamous-cell carcinomas of the lung: emerging biology, controversies, and the promise of targeted therapy. *The Lancet Oncology*, 13(10), pp.e418–426.

Duarte, I.F., Rocha, C.M., Barros, A.S., Gil, A.M., Goodfellow, B.J., Carreira, I.M., Bernardo, J., Gomes, A., Sousa, V. and Carvalho, L., 2010. Can nuclear magnetic resonance (NMR) spectroscopy reveal different metabolic signatures for lung tumours? *Virchows Archiv*, 457(6), pp.715–25.

Dunn, W.B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., Brown, M., Knowles, J.D., Halsall, A., Haselden, J.N., Nicholls, A.W., Wilson, I.D., Kell, D.B. and Goodacre, R., 2011. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature protocols*, 6(7), pp.1060–83.

Dunn, W.B., Broadhurst, D.I., Atherton, H.J., Goodacre, R. and Griffin, Julian L., 2011. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chemical Society Reviews*, 40(1), pp. 387-426.

Dwamena, B., Sonnad, S., Angobaldo, J. and Wahl, R.L., 1999. Metastases from non-small cell lung cancer: mediastinal staging in the 1990s-meta-analytic comparison of PET and CT. *Radiology*, 213(2), pp.530-6.

Eagle, H., Oyama, V.I., Levy, M., Horton, C.L. and Fleischman, R., 1956. Growth response of mammalian cells in tissue culture to L-glutamine and L-glutamic acid. *Journal of Biological Chemistry*, 218(2), pp.607–16.

Eisner, R., Stretch, C., Eastman, T., Xia, J.G., Hau, D., Damaraju, S., Greiner, R., Wishart, D.S. and Baracos, V.E., 2011. Learning to predict cancer-associated skeletal muscle wasting from H-1-NMR profiles of urinary metabolites. *Metabolomics*, 7(1), pp.25–34.

Ellis, J.K., Athersuch, T.J., Thomas, L.D.K., Teichert, F., Pérez-Trujillo, M., Svendsen, C., Spurgeon, D.J., Singh, R., Järup, L., Bundy, J.G. and Keun, H.C., 2012. Metabolic profiling detects early effects of environmental and lifestyle exposure to cadmium in a human population. *BMC Medicine*, 10(1), article ID 61, 10 pages.

Elstrom, R.L., Bauer, D.E., Buzzai, M., Karnauskas, R., Harris, M.H., Plas, D.R., Zhuang, H.M., Cinalli, R.M., Alavi, A., Rudin, C.M. and Thompson, C.B., 2004. Akt stimulates aerobic glycolysis in cancer cells. *Cancer Research*, 64(11), pp.3892–9.

Engelke, U.F.H., Sambeek, M.L.F.L., Jong, J.G.N. de, Leroy, J.G., Morava, E., Smeitink, J.A.M. and Wevers, R.A., 2004. N-Acetylated metabolites in urine: proton nuclear magnetic resonance spectroscopic study on patients with inborn errors of metabolism. *Clinical Chemistry*, 50(1), pp.58–66.

Fan, T.W.M., Lane, A.N., Higashi, R.M., Farag, M.A., Gao, H., Bousamra, M. and Miller, D.M., 2009. Altered regulation of metabolic pathways in human lung cancer discerned by 13C stable isotope-resolved metabolomics (SIRM). *Molecular Cancer*, 8(41), 19 pages.

Fang, J.S., Gillies, R.D. and Gatenby, R.A., 2008. Adaptation to hypoxia and acidosis in carcinogenesis and tumor progression. *Seminars in Cancer Biology*, 18(5), pp.330–7.

Fawcett, T., 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), pp.861–74.

Ferlay, J., Bray, F., Forman, D. and Mathers, C., 2010. GLOBOCAN 2008 v2.0, Cancer Incidence and Mortality Worldwide: IARC Cancer Base No. 10. http://globocan.iarc.fr/, accessed 3 September 2013.

Ferruzzi, E., Franceschini, R., Cazzolato, G., Geroni, C., Fowst, C., Pastorino, U., Tradati, N., Tursi, J., Dittadi, R. and Gion, M., 2003. Blood glutathione as a surrogate marker of cancer tissue glutathione S-transferase activity in non-small cell lung cancer and

squamous cell carcinoma of the head and neck. *European Journal of Cancer*, 39(7), pp.1019–29.

Fiehn, O., 2002. Metabolomics – the link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1-2), pp.155–71.

Fiorenza, A.M., Branchi, A., Cardena, A., Molgora, M., Rovellini, A. and Sommariva, D., 1996. Serum cholesterol levels in patients with cancer Relationship with nutritional status. *International Journal of Clinical and Laboratory Research*, 26(1), pp.37–42.

Fiorenza, A.M., Branchi, A. and Sommariva, D., 2000. Serum lipoprotein profile in patients with cancer. A comparison with non-cancer subjects. *International Journal of Clinical and Laboratory Research*, 30(3), pp.141–5.

Fishback, N.F., Travis, W.D., Moran, C.A., Guinee, D.G., McCarthy, W.F. and Koss, M.N., 1994. Pleomorphic (spindle/giant cell) carcinoma of the lung. A clinicopathologic correlation of 78 cases. *Cancer*, 73(12), pp.2936–45.

Folpe, A.L., Gown, A.M., Lamps, L.W., Garcia, R., Dail, D.H., Zarbo, R.J. and Schmidt, R.A., 1999. Thyroid transcription factor-1: immunohistochemical evaluation in pulmonary neuroendocrine tumors. *Modern Pathology*, 12(1), pp.5–8.

Forshed, J., Schuppe-Koistinen, I. and Jacobsson, S.P., 2003. Peak alignment of NMR signals by means of a genetic algorithm. *Analytica Chimica Acta*, 487(2), pp.189–99.

Franklin, W.A., Veve, R., Hirsch, F.R., Helfrich, B.A. and Bunn, P.A., 2002. Epidermal growth factor receptor family in lung cancer and premalignancy. *Seminars in Oncology*, 29(1), pp.3–14.

Freedman, D.S., Otvos, J.D., Jeyarajah, E.J., Shalaurova, I., Cupples, L.A., Parise, H., D'Agostino, R.B., Wilson, P.W. and Schaefer, E.J., 2004. Sex and age differences in lipoprotein subclasses measured by nuclear magnetic resonance spectroscopy: the Framingham Study. *Clinical Chemistry*, 50(7), pp.1189–200.

Früh, M., De Ruysscher, D., Popat, S., Crinò, L., Peters, S. and Felip, E., 2013. Small-cell lung cancer (SCLC): ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 24(Suppl 6), pp.vi99–105.

Funai, K., Yokose, T., Ishii, G., Araki, K., Yoshida, J., Nishimura, M., Nagai, K., Nishiwaki, Y. and Ochiai, A., 2003. Clinicopathologic characteristics of peripheral squamous cell carcinoma of the lung. *The American Journal of Surgical Pathology*, 27(7), pp.978–84.

Fyfe, C.A., 1983. *Solid state NMR for chemists*, Ontario: CFC Press.

Gadducci, A., Brunetti, I., Muttini, M.P., Fanucchi, A., Dargenio, F., Giannessi, P.G. and Conte, P.F., 1994. Epidoxorubicin and lonidamine in refractory or recurrent epithelial ovarian cancer. *European Journal of Cancer*, 30A(10), pp.1432–5.

Gamcsik, M.P., Kasibhatla, M.S., Teeter, S.D. and Colvin, O.M., 2012. Glutathione levels in human tumors. *Biomarkers*, 17(8), pp.671–91.

Gao, H., Dong, B., Liu, X., Xuan, H., Huang, Y. and Lin, D., 2008. Metabonomic profiling of renal cell carcinoma: High-resolution proton nuclear magnetic resonance spectroscopy of human serum with multivariate data analysis. *Analytica Chimica Acta*, 624(2), pp.269–77.

Gao, H.C., Lu, Q., Liu, X., Cong, H., Zhao, L.C., Wang, H.M. and Lin, D.H., 2009. Application of H-1 NMR-based metabonomics in the study of metabolic profiling of human hepatocellular carcinoma and liver cirrhosis. *Cancer Science*, 100(4), pp.782–5.

Gao, P., Tchernyshyov, I., Chang, T.C., Lee, Y.S., Kita, K., Ochi, T., Zeller, K.I., De Marzo, A.M., Van Eyk, J.E., Mendell, J.T. and Dang, C. V, 2009. c-Myc suppression of miR-23a/b enhances mitochondrial glutaminase expression and glutamine metabolism. *Nature*, 458(7239), pp.762–5.

Garcia, A.A., Leichman, L., Baranda, J., Pandit, L., Lenz, H.-J. and Leichman, C.G., 2003. Phase II clinical trial of 5-fluorouracil, trimetrexate, and leucovorin (NFL) in patients with advanced pancreatic cancer. *International Journal of Gastrointestinal Cancer*, 34(2-3), pp.79–86.

Garrod, S., Humpfer, E., Spraul, M., Connor, S.C., Polley, S., Connelly, J., Lindon, J.C., Nicholson, J.K. and Holmes, E., 1999. High-resolution magic angle spinning 1H NMR spectroscopic studies on intact rat renal cortex and medulla. *Magnetic Resonance in Medicine*, 41(6), pp.1108–18.

Gatenby, R.A. and Gillies, R.J., 2007. Glycolysis in cancer: a potential target for therapy. *Internation Journal of Biochemistry & Cell Biology*, 39(7-8), pp.1358–66.

Gatzemeier, U., Cavalli, F., Häussinger, K., Kaukel, E., Koschel, G., Martinelli, G., Neuhauss, R. and von Pawel, J., 1991. Phase III trial with and without lonidamine in non-small cell lung cancer. *Seminars in oncology*, 18(2 Suppl 4), pp.42–8.

De Geus-Oei, L.-F., van Krieken, J.H.J.M., Aliredjo, R.P., Krabbe, P.F.M., Frielink, C., Verhagen, A.F.T., Boerman, O.C. and Oyen, W.J.G., 2007. Biological correlates of FDG uptake in non-small cell lung cancer. *Lung Cancer*, 55(1), pp.79–87.

Gika, H.G., Theodoridis, G.A. and Wilson, I.D., 2008. Hydrophilic interaction and reversed-phase ultra-performance liquid chromatography TOF-MS for metabonomic analysis of Zucker rat urine. *Journal of Separation Science*, 31(9), pp.1598–608.

Glaudemans, A.W.J.M., Enting, R.H., Heesters, M.A.A.M., Dierckx, R.A.J.O., van Rheenen, R.W.J., Walenkamp, A.M.E. and Slart, R.H.J.A., 2013. Value of 11C-methionine PET in imaging brain tumours and metastases. *European Journal of Nuclear Medicine and Molecular Imaging*, 40(4), pp.615–35.

Golman, K., Olsson, L.E., Axelsson, O., Månsson, S., Karlsson, M. and Petersson, J.S., 2003. Molecular imaging using hyperpolarized 13C. *The British Journal of Radiology*, 76(2), pp.S118–27.

Gordan, J.D. and Simon, M.C., 2007. Hypoxia-inducible factors: central regulators of the tumor phenotype. *Current Opinion in Genetics & Development*, 17(1), pp.71–7.

Gottlieb, E. and Vousden, K.H., 2010. p53 regulation of metabolic pathways. *Cold Spring Harbor Perspectives in Biology*, 2(4), article D a001040, 11 pages.

Gould, M.K., Donington, J., Lynch, W.R., Mazzone, P.J., Midthun, D.E., Naidich, D.P. and Wiener, R.S., 2013. Evaluation of individuals with pulmonary nodules: when is it lung cancer? Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest*, 143(Suppl 5), pp.S93–120.

Gould, M.K., Maclean, C.C., Kuschner, W.G., Rydzak, C.E. and Owens, D.K., 2001. Accuracy of positron emission tomography for diagnosis of pulmonary nodules and mass lesions: a meta-analysis. *Journal of the American Medical Association*, 285(7), pp.914–24.

Graça, G., Moreira, A.S., Correia, A.J. V, Goodfellow, B.J., Barros, A.S., Duarte, I.F., Carreira, I.M., Galhano, E., Pita, C., Almeida, M. do C. and Gil, A.M., 2013. Mid-infrared (MIR) metabolic fingerprinting of amniotic fluid: a possible avenue for early diagnosis of prenatal disorders? *Analytica Chimica Acta*, 764, pp.24–31.

Grassi, I., Nanni, C., Allegri, V., Morigi, J.J., Montini, G.C., Castellucci, P. and Fanti, S., 2012. The clinical use of PET with (11)C-acetate. *American Journal of Nuclear Medicine and Molecular Imaging*, 2(1), pp.33–47.

Griffin, J.L. and Kauppinen, R.A., 2007. A metabolomics perspective of human brain tumours. *FEBS Journal*, 274(5), pp.1132–9.

Guinee, D.G., Fishback, N.F., Koss, M.N., Abbondanzo, S.L. and Travis, W.D., 1994. The spectrum of immunohistochemical staining of small-cell lung carcinoma in specimens from transbronchial and open-lung biopsies. *American Journal of Clinical Pathology*, 102(4), pp.406–14.

Günther, H., 2013. *NMR spectroscopy: basic principles, concepts, and applications in chemistry* 3rd Edition, Weinheim: John Wiley & Sons, Ltd.

Guo, J.F., Higashi, K., Yokota, H., Nagao, Y., Ueda, Y., Kodama, Y., Oguchi, M., Taki, S., Tonami, H. and Yamamoto, I., 2004. In vitro proton magnetic resonance spectroscopic lactate and choline measurements, F-18-FDG uptake, and prognosis in patients with lung adenocarcinoma. *Journal of Nuclear Medicine*, 45(8), pp.1334–9.

Guo, Y.M., Wang, X.M., Qiu, L., Qin, X.Z., Liu, H., Wang, Y.Y., Li, F., Wang, X.D., Chen, G.Q., Song, G.G., Li, F.J., Guo, S. and Li, Z.L., 2012. Probing gender-specific lipid metabolites and diagnostic biomarkers for lung cancer using Fourier transform ion cyclotron resonance mass spectrometry. *Clinica Chimica Acta*, 414, pp.135–41.

Haffner, H.T., Graw, M., Besserer, K., Blickle, U. and Henge, C., 1996. Endogenous methanol: variability in concentration and rate of production. Evidence of a deep compartment? *Forensic Science International*, 79(2), pp.145–54.

Hanahan, D. and Weinberg, R.A., 2011. Hallmarks of cancer: the next generation. *Cell*, 144(5), pp.646–74.

Hanahan, D. and Weinberg, R.A., 2000. The hallmarks of cancer. *Cell*, 100(1), pp.57–70.

Hanaoka, H., Yoshioka, Y., Ito, I., Niitu, K. and Yasuda, N., 1993. In vitro characterization of lung cancers by the use of 1H nuclear magnetic resonance spectroscopy of tissue extracts and discriminant factor analysis. *Magnetic Resonance in Medicine*, 29(4), pp.436–40.

Harris, R., Patel, S.U., Sadler, P.J. and Viles, J.H., 1996. Observation of albumin resonances in proton nuclear magnetic resonance spectra of human blood plasma: N-

terminal assignments aided by use of modified recombinant albumin. *The Analyst*, 121(7), pp.913–22.

Hasim, A., Ma, H., Mamtimin, B., Abudula, A., Niyaz, M., Zhang, L.W., Anwer, J. and Sheyhidin, I., 2012. Revealing the metabonomic variation of EC using (1)H-NMR spectroscopy and its association with the clinicopathological characteristics. *Molecular Biology Reports*, 39(9), pp.8955–64.

Hassanein, M., Callison, J.C., Callaway-Lane, C., Aldrich, M.C., Grogan, E.L. and Massion, P.P., 2012. The state of molecular biomarkers for the early detection of lung cancer. *Cancer Prevention Research*, 5(8), pp.992–1006.

Hatzivassiliou, G., Zhao, F.P., Bauer, D.E., Andreadis, C., Shaw, A.N., Dhanak, D., Hingorani, S.R., Tuveson, D.A. and Thompson, C.B., 2005. ATP citrate lyase inhibition can suppress tumor cell growth. *Cancer Cell*, 8(4), pp.311–21.

Hausmann, D., Bittencourt, L.K., Attenberger, U.I., Sertdemir, M., Weidner, A., Büsing, K.A., Brade, J., Wenz, F., Schoenberg, S.O. and Dinter, D.J., 2014. Diagnostic accuracy of 18F choline PET/CT using time-of-flight reconstruction algorithm in prostate cancer patients with biochemical recurrence. *Clinical Nuclear Medicine*, 39(3), pp.197–201.

Hecht, S.S., 2012. Lung carcinogenesis by tobacco smoke. *International Journal of Cancer*, 131(12), pp.2724–32.

Heiden, M.G. V, Cantley, L.C., Thompson, C.B. and Vander Heiden, M.G., 2009. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science*, 324(5930), pp.1029–33.

Heinrich, P.C., Morris, H.P. and Weber, G., 1976. Behavior of transaldolase (EC-2.2.1.2) and transketolase (EC-2.2.1.1) activities in normal, neoplastic, differentiating, and regenerating liver. *Cancer Research*, 36(9), pp.3189–97.

Heinzmann, S.S., Merrifield, C.A., Rezzi, S., Kochhar, S., Lindon, J.C., Holmes, E. and Nicholson, J.K., 2012. Stability and robustness of human metabolic phenotypes in response to sequential food challenges. *Journal of Proteome Research*, 11(2), pp.643–55.

Heo, S.H., Lee, S.J., Ryoo, H.M., Park, J.Y. and Cho, J.Y., 2007. Identification of putative serum glycoprotein biomarkers for human lung adenocarcinoma by multilectin affinity chromatography and LC-MS/MS. *Proteomics*, 7(23), pp.4292–302.

Hespanhol, V., Parente, B., Araújo, A., Cunha, J., Fernandes, A., Figueiredo, M.M., Neveda, R., Soares, M., João, F. and Queiroga, H., 2013. Lung cancer in Northern Portugal: A hospital-based study. *Revista Portuguesa de Pneumologia*, 19(6), pp.245–51.

Hipkiss, A.R., 2010. Aging, proteotoxicity, mitochondria, glycation, NAD and carnosine: possible inter-relationships and resolution of the oxygen paradox. *Frontiers in Aging Neuroscience*, 2, article ID 10, 6 pages.

Hocker, J.R., Peyton, M.D., Lerner, M.R., Mitchell, S.L., Lightfoot, S.A., Lander, T.J., Bates-Albers, L.M., Vu, N.T., Hanas, R.J., Kupiec, T.C., Brackett, D.J. and Hanas, J.S., 2011. Serum discrimination of early-stage lung cancer patients using electrospray-ionization mass spectrometry. *Lung Cancer*, 74(2), pp.206–11.

Hoffmann, E. de and Stroobant, V., 2007. *Mass Spectrometry: Principals and Applications* 3rd Edition, Chichester, UK: John Wiley & Sons, Ltd.

Holmes, E., Foxall, P.J.., Spraul, M., Duncan Farrant, R., Nicholson, J.K. and Lindon, J.C., 1997. 750 MHz 1H NMR spectroscopy characterisation of the complex metabolic pattern of urine from patients with inborn errors of metabolism: 2-hydroxyglutaric aciduria and maple syrup urine disease. *Journal of Pharmaceutical and Biomedical Analysis*, 15(11), pp.1647–59.

Holmes, E., Li, J. V, Athanasiou, T., Ashrafian, H. and Nicholson, J.K., 2011. Understanding the role of gut microbiome–host metabolic signal disruption in health and disease. *Trends in Microbiology*, 19(7), pp.349–59.

Hori, S., Nishiumi, S., Kobayashi, K., Shinohara, M., Hatakeyama, Y., Kotani, Y., Hatano, N., Maniwa, Y., Nishio, W., Bamba, T., Fukusaki, E., Azuma, T., Takenawa, T., Nishimura, Y. and Yoshida, M., 2011. A metabolomic approach to lung cancer. *Lung Cancer*, 74(2), pp.284–92.

Howlader, N., Noone, A., Krapcho, M., Garshell, J., Miller, D., Altekruse, S., Kosary, C., Yu, M., Ruhl, J., Tatalovich, Z., Mariotto, A., Lewis, D., Chen, H., Feuer, E. and Cronin, K., 2014. SEER cancer statistics review (CSR) 1975-2011. *National Cancer Institute*. http://seer.cancer.gov/csr/1975_2011/, accessed 1 May 2014.

Hu, W., Zhang, C., Wu, R., Sun, Y., Levine, A. and Feng, Z., 2010. Glutaminase 2, a novel p53 target gene regulating energy metabolism and antioxidant function. *Proceedings*

of the National Academy of Sciences of the United States of America, 107(16), pp.7455–60.

Idowu, M.O. and Powers, C.N., 2010. Lung cancer cytology: potential pitfalls and mimics - a review. *International Journal of Clinical and Experimental Pathology*, 3(4), pp.367–85.

Ilonen, I.K., Räsänen, J. V, Sihvo, E.I., Knuuttila, A., Salmenkivi, K.M., Ahotupa, M.O., Kinnula, V.L. and Salo, J.A., 2009. Oxidative stress in non-small cell lung cancer: role of nicotinamide adenine dinucleotide phosphate oxidase and glutathione. *Acta Oncologica*, 48(7), pp.1054–61.

Imperiale, A., Elbayed, K., Moussallieh, F-M., Neuville, A., Piotto, M., Bellocq, J-P., Lutz, P., Namer, I-J., 2011. Metabolomic pattern ofchildhood neuroblastoma obtained by 1H-high-resolution magic angle spinning (HRMAS) NMR spectroscopy. *Pediatric Blood & Cancer*, 56(1), pp.24-34.

Ishida, T., Kaneko, S., Yokoyama, H., Inoue, T., Sugio, K. and Sugimachi, K., 1992. Adenosquamous carcinoma of the lung. Clinicopathologic and immunohistochemical features. *American Journal of Clinical Pathology*, 97(5), pp.678–85.

Ito, H., Matsuo, K., Tanaka, H., Koestler, D.C., Ombao, H., Fulton, J., Shibata, A., Fujita, M., Sugiyama, H., Soda, M., Sobue, T. and Mor, V., 2011. Nonfilter and filter cigarette consumption and the incidence of lung cancer by histological type in Japan and the United States: analysis of 30-year data from population-based cancer registries. *International Journal of Cancer*, 128(8), pp.1918–28.

Iyoda, A., Hiroshima, K., Toyozaki, T., Haga, Y., Fujisawa, T. and Ohwada, H., 2001. Clinical characterization of pulmonary large cell neuroendocrine carcinoma and large cell carcinoma with neuroendocrine morphology. *Cancer*, 91(11), pp.1992–2000.

Jacobsen, N.E., 2007. *NMR spectroscopy explained: simplified theory, applications and examples for organic chemistry and structural biology*, 1st Edition. New Jersey: John Wiley & Sons, Inc.

Jain, M., Nilsson, R., Sharma, S., Madhusudhan, N., Kitami, T., Souza, A.L., Kafri, R., Kirschner, M.W., Clish, C.B. and Mootha, V.K., 2012. Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. *Science*, 336(6084), pp.1040–4.

Jiang, S.X., Kameya, T., Shoji, M., Dobashi, Y., Shinada, J. and Yoshimura, H., 1998. Large cell neuroendocrine carcinoma of the lung: a histologic and immunohistochemical study of 22 cases. *The American Journal of Surgical Pathology*, 22(5), pp.526–37.

Jones, N.P. and Schulze, A., 2012. Targeting cancer metabolism – aiming at a tumour's sweet-spot. *Drug Discovery Today*, 17(5-6), pp.232–41.

Jong, C., Azuma, J., Schaffer, S., 2012. Mechanism underlying the antioxidant activity of taurine: prevention of mitochondrial oxidant production. *Amino Acids*, 42(6), pp.2223-32.

Jordan, K.W., Adkins, C.B., Su, L., Halpern, E.F., Mark, E.J., Christiani, D.C. and Cheng, L.L., 2010. Comparison of squamous cell carcinoma and adenocarcinoma of the lung by metabolomic analysis of tissue–serum pairs. *Lung Cancer*, 68(1), pp.44–50.

Joseph, J., Cardesa, A. and Carreras, J., 1997. Creatine kinase activity and isoenzymes in lung, colon and liver carcinomas. *British Journal of Cancer*, 76(5), pp.600–5.

Kami, K., Fujimori, T., Sato, H., Sato, M., Yamamoto, H., Ohashi, Y., Sugiyama, N., Ishihama, Y., Onozuka, H., Ochiai, A., Esumi, H., Soga, T. and Tomita, M., 2013. Metabolomic profiling of lung and prostate tumor tissues by capillary electrophoresis time-of-flight mass spectrometry. *Metabolomics*, 9(2), pp.444–53.

Kassel, D.B., Martin, M., Schall, W. and Sweeley, C.C., 1986. Urinary metabolites of L-threonine in type 1 diabetes determined by combined gas chromatography/chemical ionization mass spectrometry. *Biomedical & Environmental Mass Spectrometry*, 13(10), pp.535–40.

Katajamaa, M., Miettinen, J. and Oresic, M., 2006. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics*, 22(5), pp.634–6.

Kaufman, O. and Dietel, M., 2000. Expression of thyroid transcription factor-1 in pulmonary and extrapulmonary small cell carcinomas and other neuroendocrine carcinomas of various primary sites. *Histopathology*, 36(5), pp.415–20.

Kaur, P., Rizk, N., Ibrahim, S., Luo, Y., Younes, N., Perry, B., Dennis, K., Zirie, M., Luta, G. and Cheema, A.K., 2013. Quantitative metabolomic and lipidomic profiling reveals aberrant amino acid metabolism in type 2 diabetes. *Molecular Biosystems*, 9(2), pp.307–17.

Keeler, J., 2010. *Understanding NMR spectroscopy*, 1st Edition. Wiley.

Kerr, K., 2012. Personalized medicine for lung cancer: new challenges for pathology. *Histopathology*, 60(4), pp.531-46.

Keun, H.C., Athersuch, T.J., Beckonert, O., Wang, Y., Saric, J., Shockcor, J.P., Lindon, J.C., Wilson, I.D., Holmes, E. and Nicholson, J.K., 2008. Heteronuclear 19F-1H statistical total correlation spectroscopy as a tool in drug metabolism: study of flucloxacillin biotransformation. *Analytical Chemistry*, 80(4), pp.1073–9.

Khuder, S.A., 2001. Effect of cigarette smoking on major histological types of lung cancer: a meta-analysis. *Lung Cancer*, 31(2-3), pp.139–48.

Kim, E.H. and E.Misek, D., 2011. Glycoproteomics-based identification of cancer biomarkers. *International Journal of Proteomics*, 2011, article ID 601937, 10 pages.

Kim, J.W., Tchernyshyov, I., Semenza, G.L. and Dang, C. V, 2006. HIF-1-mediated expression of pyruvate dehydrogenase kinase: A metabolic switch required for cellular adaptation to hypoxia. *Cell Metabolism*, 3(3), pp.177–85.

Kimmelman, A.C., 2011. The dynamic nature of autophagy in cancer. *Genes & Development*, 25(19), pp.1999–2010.

Kochhar, S., Jacobs, D.M., Ramadan, Z., Berruex, F., Fuerhoz, A. and Fay, L.B., 2006. Probing gender-specific metabolism differences in humans by nuclear magnetic resonance-based metabonomics. *Analytical Biochemistry*, 352(2), pp.274–81.

Kohl, S.M., Klein, M.S., Hochrein, J., Oefner, P.J., Spang, R. and Gronwald, W., 2012. State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics*, 8(Suppl 1), pp.S146–60.

Kuhajda, F.P., Jenner, K., Wood, F.D., Hennigar, R.A., Jacobs, L.B., Dick, J.D. and Pasternack, G.R., 1994. Fatty acid synthesis - A potential selective target for antieoplastic therapy. *Proceedings of the National Academy of Sciences of the United States of America*, 91(14), pp.6379–83.

Kumps, A., Duez, P. and Mardens, Y., 2002. Metabolic, nutritional, iatrogenic and artifactual sources of urinary organic acids: a comprehensive table. *Clinical Chemistry*, 48(5), pp.708–17.

Kurhanewicz, J., Vigneron, D.B., Brindle, K., Chekmenev, E.Y., Comment, A., Cunningham, C.H., Deberardinis, R.J., Green, G.G., Leach, M.O., Rajan, S.S., Rizi, R.R., Ross, B.D., Warren, W.S. and Malloy, C.R., 2011. Analysis of cancer

metabolism by imaging hyperpolarized nuclei: prospects for translation to clinical research. *Neoplasia*, 13(2), pp.81–97.

Lababede, O., Meziane, M. and Rice, T., 2011. Seventh edition of the cancer staging manual and stage grouping of lung cancer: quick reference chart and diagrams. *Chest*, 139(1), pp.183–9.

Lai, H.-S., Lee, J.-C., Lee, P.-H., Wang, S.-T. and Chen, W.-J., 2005. Plasma free amino acid profile in cancer patients. *Seminars in Cancer Biology*, 15(4), pp.267–76.

Lau, S.K., Luthringer, D.J. and Eisen, R.N., 2002. Thyroid transcription factor-1: a review. *Applied Immunohistochemistry & Molecular Morphology*, 10(2), pp.97–102.

Lawton, K.A., Berger, A., Mitchell, M., Milgram, K.E., Evans, A.M., Guo, L., Hanson, R.W., Kalhan, S.C., Ryals, J.A. and Milburn, M. V, 2008. Analysis of the adult human plasma metabolome. *Pharmacogenomics*, 9(4), pp.383–97.

Leij-Halfwerk, S., Dagnelie, P.C., van den Berg, J.W.O., Wattimena, J.D.L., Hordijk-Luijk, C.H. and Wilson, J.H.P., 2000. Weight loss and elevated gluconeogenesis from alanine in lung cancer patients. *American Journal of Clinical Nutrition*, 71(2), pp.583–9.

De Lena, M., Lorusso, V., Latorre, A., Fanizza, G., Gargano, G., Caporusso, L., Guida, M., Catino, A., Crucitta, E., Sambiasi, D. and Mazzei, A., 2001. Paclitaxel, cisplatin and lonidamine in advanced ovarian cancer. A phase II study. *European Journal of Cancer*, 37(3), pp.364–8.

Larsen, J. and Minna, J., 2011. Molecular biology of lung cancer: clinical implications. *Clinics in Chest Medicine*, 32(4), pp.703-40.

Lenz, E.M. and Wilson, I.D., 2007. Analytical strategies in metabonomics. *Journal of Proteome Research*, 6(2), pp.443–58.

Li, M., Peng, Z., Liu, Q., Sun, J., Yao, S. and Liu, Q., 2013. Value of 11C-choline PET/CT for lung cancer diagnosis and the relation between choline metabolism and proliferation of cancer cells. *Oncology Reports*, 29(1), pp.205–11.

Lieberman, B.P., Ploessl, K., Wang, L., Qu, W., Zha, Z., Wise, D.R., Chodosh, L.A., Belka, G., Thompson, C.B. and Kung, H.F., 2011. PET imaging of glutaminolysis in tumors by 18F-(2S,4R)4-fluoroglutamine. *Journal of Nuclear Medicine*, 52(12), pp.1947–55.

Lin, X., Wang, Q., Yin, P., Tang, L., Tan, Y., Li, H., Yan, K. and Xu, G., 2011. A method for handling metabonomics data from liquid chromatography/mass spectrometry: combinational use of support vector machine recursive feature elimination, genetic algorithm and random forest for feature selection. *Metabolomics*, 7(4), pp.549–58.

Lindon, J.C., Holmes, E. and Nicholson, J.K., 2007. Metabonomics in pharmaceutical R&D. *The FEBS Journal*, 274(5), pp.1140–51.

Liu, H., Wang, H., Li, C., Wang, L., Pan, Z. and Wang, L., 2014. Investigation of volatile organic metabolites in lung cancer pleural effusions by solid-phase microextraction and gas chromatography/mass spectrometry. *Journal of Chromatography B*, 945-946, pp.53–9.

Liu, M., Tang, H., Nicholson, J.K. and Lindon, J.C., 2002. Use of 1H NMR-determined diffusion coefficients to characterize lipoprotein fractions in human blood plasma. *Magnetic Resonance in Chemistry*, 40(13), pp.S83–8.

Lokhov, P.G., Kharybin, O.N. and Archakov, A.I., 2012. Diagnosis of lung cancer based on direct-infusion electrospray mass spectrometry of blood plasma metabolites. *International Journal Mass Spectrometry*, 309, pp.200–5.

Lokhov, P.G., Trifonova, O.P., Maslov, D.L. and Archakov, A.I., 2013. Blood plasma metabolites and the risk of developing lung cancer in Russia. *European Journal of Cancer Prevention*, 22(4), pp.335–41.

Lowe, I., 1959. Free Induction Decays of Rotating Solids. *Physical Review Letters*, 2(7), pp.285–287.

Ludwig, C. and Viant, M.R., 2010. Two-dimensional J-resolved NMR spectroscopy: review of a key methodology in the metabolomics toolbox. *Phytochemical Analysis*, 21(1), pp.22–32.

Ma, Y.L., Qin, H.L., Liu, W.J., Peng, J.Y., Huang, L., Zhao, X.P. and Cheng, Y.Y., 2009. Ultra-high performanceliquid chromatography-mass spectrometry for the metabolomic analysis of urine in colorectal cancer. *Digestive Diseases and Sciences*, 54(12), pp.2655–62.

Macheda, M.L., Rogers, S. and Best, J.D., 2005. Molecular and cellular regulation of glucose transporter (GLUT) proteins in cancer. *Journal of Cellular Physiology*, 202(3), pp.654–62.

MacIntyre, D.A., Jiménez, B., Lewintre, E.J., Martín, C.R., Schäfer, H., MBallesteros, C.G., Mayans, J.R., Spraul, M., García-Conde, J. and Pineda-Lucena, A., 2010. Serum metabolome analysis by 1H-NMR revelas differences between chronic lymphocytic leukaemia molecular subgroups. *Leukemia*, 24(4), pp.788–97.

Maddocks, O.D.K., Berkers, C.R., Mason, S.M., Zheng, L., Blyth, K., Gottlieb, E. and Vousden, K.H., 2013. Serine starvation induces stress and p53-dependent metabolic remodelling in cancer cells. *Nature*, 493(7433), pp.542–6.

Madhok, B.M., Yeluri, S., Perry, S.L., Hughes, T.A. and Jayne, D.G., 2011. Targeting glucose metabolism: an emerging concept for anticancer therapy. *American Journal of Clinical Oncology*, 34(6), pp.628–35.

Maeda, J., Higashiyama, M., Imaizumi, A., Nakayama, T., Yamamoto, H., Daimon, T., Yamakado, M., Imamura, F. and Kodama, K., 2010. Possibility of multivariate function composed of plasma amino acid profiles as a novel screening index for non-small cell lung cancer: a case control study. *BMC Cancer*, 10(690), article ID 690, 8 pages.

Maher, A.D., Cysique, L.A., Brew, B.J. and Rae, C.D., 2011. Statistical integration of 1H NMR and MRS data from different biofluids and tissues enhances recovery of biological information from individuals with HIV-1 infection. *Journal of Proteome Research*, 10(4), pp.1737–45.

Mallol, R., Rodriguez, M.A., Brezmes, J., Masana, L. and Correig, X., 2013. Human serum/plasma lipoprotein analysis by NMR: application to the study of diabetic dyslipidemia. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 70, pp.1–24.

Marín-Hernández, A., Gallardo-Pérez, J.C., Ralph, S.J., Rodríguez-Enríquez, S. and Moreno-Sánchez, R., 2009. HIF-1alpha modulates energy metabolism in cancer cells by inducing over-expression of specific glycolytic isoforms. *Mini Reviews in Medicinal Chemistry*, 9(9), pp.1084–101.

Martinez-Granados, B., Monleon, D., Martinez-Bisbal, M.C., Rodrigo, J.M., Olmo, J. del, Lluch, P., Fernandez, A., Marti-Bonmati, L. and Celda, B., 2006. Metabolite identification in human liver needle biopsies by high-resolution magic angle spinning 1H NMR spectroscopy. *NMR in Biomedicine*, 19(1), pp.90–100.

Marzocco, S., Di Paola, R., Ribecco, M.T., Sorrentino, R., Domenico, B., Genesio, M., Pinto, A., Autore, G. and Cuzzocrea, S., 2004. Effect of methylguanidine in a model of septic shock induced by LPS. *Free Radical Research*, 38(11), pp.1143-53.

Mathe, E.A., Patterson, A.D., Haznadar, M., Manna, S.K., Krausz, K.W., Bowman, E.D., Shields, P.G., Idle, J.R., Smith, P.B., Anami, K., Kazandjian, D., Hatzakis, E., Gonzalez, F.J. and Harris, C.C., 2014. Non-invasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer. *Cancer Research*, 74(12), pp.3259–70.

Mazurek, S., 2007. Tumor cell energetic metabolome. In V. Saks, ed. *Molecular System Bioenergetics*. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, pp. 521–40.

Mazurek, S., Boschek, C.B., Hugoc, F. and Eigenbrodt, E., 2005. Pyruvate kinase type M2 and its role in tumor growth and spreading. *Seminars in Cancer Biology*, 15(4), pp.300–8.

McClay, J.L., Adkins, D.E., Isern, N.G., O'Connell, T.M., Wooten, J.B., Zedler, B.K., Dasika, M.S., Webb, B.T., Webb-Robertson, B.J., Pounds, J.G., Murrelle, E.L., Leppert, M.F. and van den Oord, E., 2010. H-1 nuclear magnetic resonance metabolomics analysis identifies novel urinary biomarkers for lung function. *Journal of Proteome Research*, 9(6), pp.3083–90.

Meiboom, S. and Gill, D., 1958. Modified spin-echo method for measuring Nuclear relaxation times. *The Review of Scientific Intruments*, 29(8), pp.688–91.

Meijer, T.W.H., Schuurbiers, O.C.J., Kaanders, J.H.A.M., Looijen-Salamon, M.G., de Geus-Oei, L.-F., Verhagen, A.F.T.M., Lok, J., van der Heijden, H.F.M., Rademakers, S.E., Span, P.N. and Bussink, J., 2012. Differences in metabolism between adeno- and squamous cell non-small cell lung carcinomas: spatial distribution and prognostic value of GLUT1 and MCT4. *Lung Cancer*, 76(3), pp.316–23.

Mirsadraee, S., Oswal, D., Alizadeh, Y., Caulo, A. and van Beek, E., 2012. The 7th lung cancer TNM classification and staging system: Review of the changes and implications. *World Journal of Radiology*, 4(4), pp.128–34.

Miyagi, Y., Higashiyama, M., Gochi, A., Akaike, M., Ishikawa, T., Miura, T., Saruki, N., Bando, E., Kimura, H., Imamura, F., Moriyama, M., Ikeda, I., Chiba, A., Oshita, F., Imaizumi, A., Yamamoto, H., Miyano, H., Horimoto, K., Tochikubo, O.,

Mitsushima, T., Yamakado, M. and Okamoto, N., 2011. Plasma free amino acid profiling of five types of cancer patients and its application for early detection. *PLoS ONE*, 6(9), article ID e24143, 12 pages.

Moestue, S., Sitter, B., Bathen, T.F., Tessem, M.-B.B. and Gribbestad, I.S., 2011. HR MAS MR spectroscopy in metabolic characterization of cancer. *Current Topics in Medicinal Chemistry*, 11(1), pp.2–26.

Mohamed, A., Deng, X., Khuri, F.R. and Owonikoko, T.K., 2014. Altered glutamine metabolism and therapeutic opportunities for lung cancer. *Clinical lung cancer*, 15(1), pp.7–15.

Mohanti, B.K., Rath, G.K., Anantha, N., Kannan, V., Das, B.S., Chandramouli, B.A.R., Banerjee, A.K., Das, S., Jena, A., Ravichandran, R., Sahi, U.P., Kumar, R., Kapoor, N., Kalia, V.K., Dwarakanath, B.S. and Jain, V., 1996. Improving cancer radiotherapy with 2-deoxy-d-glucose: phase I/II clinical trials on human cerebral gliomas. *International Journal of Radiation Oncology*, 35(1), pp.103–11.

Molina, A.R. de, Rodríguez-González, A., Gutiérrez, R., Martínez-Piñeiro, L., Sánchez, J.J., Bonilla, F., Rosell, R. and Lacal, J.C., 2002. Overexpression of choline kinase is a frequent feature in human tumor-derived cell lines and in lung, prostate, and colorectal human cancers. *Biochemical and Biophysical Research Communications*, 296(3), pp.580–3.

Molina, A.R. de, Sarmentero-Estrada, J., Belda-Iniesta, C., Tarón, M., Molina, V.R. de, Cejas, P., Skrzypski, M., Gallego-Ortega, D., Castro, J. de, Casado, E., García-Cabezas, M.A., Sánchez, J.J., Nistal, M., Rosell, R., González-Barón, M. and Lacal, J.C., 2007. Expression of choline kinase alpha to predict outcome in patients with early-stage non-small-cell lung cancer: a retrospective study. *Lancet Oncology*, 8(10), pp.889–97.

Morrone, F.B., Jacques-Silva, M.C., Horn, A.P., Bernardi, A., Schwartsmann, G., Rodnight, R. and Lenz, G., 2003. Extracellular nucleotides and nucleosides induce proliferation and increase nucleoside transport in human glioma cell lines. *Journal of Neuro-oncology*, 64(3), pp.211–8.

Mueller, C., Al-Batran, S., Jaeger, E., Schmidt, B., Bausch, M., Unger, C. and Sethuraman, N., 2008. A phase IIa study of PEGylated glutaminase (PEG-PGA) plus 6-diazo-5-

oxo-L-norleucine (DON) in patients with advanced refractory solid tumors. *ASCO Meeting Abstracts*, 26(Suppl 15), article ID 2533.

Muntoni, S., Atzori, L., Mereu, R., Satta, G., Macis, M.D., Congia, M., Tedde, A. and Desogus, A., 2009. Serum lipoproteins and cancer. *Nutrition, Metabolism and Cardiovascular Diseases*, 19(3), pp.218–25.

Nahon, P., Amathieu, R., Triba, M.N., Bouchemal, N., Nault, J.-C., Ziol, M., Seror, O., Dhonneur, G., Trinchet, J.-C., Beaugrand, M. and Le Moyec, L., 2012. Identification of serum proton NMR metabolomic fingerprints associated with hepatocellular carcinoma in patients with alcoholic cirrhosis. *Clinical Cancer Research*, 18(24), pp.6714–22.

Nakagawa, S. and Cuthill, I.C., 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biology Review of the Cambridge Philosophy Society*, 82(4), pp.591–605.

Nakajima, M., Kasai, T., Hashimoto, H., Iwata, Y. and Manabe, H., 1999. Sarcomatoid carcinoma of the lung: a clinicopathologic study of 37 cases. *Cancer*, 86(4), pp.608–16.

Napoli, C., Sperandio, N., Lawlor, R.T., Scarpa, A., Molinari, H. and Assfalg, M., 2012. Urine metabolic signature of pancreatic ductal adenocarcinoma by (1)h nuclear magnetic resonance: identification, mapping, and evolution. *Journal of Proteome Research*, 11(2), pp.1274–83.

Nelson, D.L., Cox, M.M., 2004. *Lehninger Principles of Biochemistry*, 4th edition. New York: Freeman, W. H. & Company.

Nelson, S.J., Kurhanewicz, J., Vigneron, D.B., Larson, P.E.Z., Harzstark, A.L., Ferrone, M., van Criekinge, M., Chang, J.W., Bok, R., Park, I., Reed, G., Carvajal, L., Small, E.J., Munster, P., Weinberg, V.K., Ardenkjaer-Larsen, J.H., Chen, A.P., Hurd, R.E., Odegardstuen, L.-I., Robb, F.J., Tropp, J. and Murray, J.A., 2013. Metabolic imaging of patients with prostate cancer using hyperpolarized [1-^{13}C]pyruvate. *Science Translational Medicine*, 5(198), article ID 198ra108, 22 pages.

New, L.-S. and Chan, E.C.Y., 2008. Evaluation of BEH C18, BEH HILIC, and HSS T3 (C18) column chemistries for the UPLC-MS-MS analysis of glutathione, glutathione disulfide, and ophthalmic acid in mouse liver and human plasma. *Journal of Chromatographic Science*, 46(3), pp.209–14.

Nicholson, J.K., Foxall, P.J.D., Spraul, M., Farrant, R.D. and Lindon, J.C., 1995. 750-MHZ H-1 and H-1-C-13 NMR-spectroscopy of human blood plasma. *Analytical Chemistry*, 67(5), pp.793–811.

Nicholson, J.K., Lindon, J.C. and Holmes, E., 1999. "Metabonomics": understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, 29(11), pp.1181–9.

Nicholson, S.A., Beasley, M.B., Brambilla, E., Hasleton, P.S., Colby, T. V, Sheppard, M.N., Falk, R. and Travis, W.D., 2002. Small cell lung carcinoma (SCLC): a clinicopathologic study of 100 cases with surgical specimens. *The American Journal of Surgical Pathology*, 26(9), pp.1184–97.

O'Connell, M.J., Sargent, D.J., Windschitl, H.E., Shepherd, L., Mahoney, M.R., Krook, J.E., Rayson, S., Morton, R.F., Rowland, K.M. and Kugler, J.W., 2006. Randomized clinical trial of high-dose levamisole combined with 5-fluorouracil and leucovorin as surgical adjuvant therapy for high-risk colon cancer. *Clinical Colorectal Cancer*, 6(2), pp.133–9.

Okamoto, N., 2012. Use of "AminoIndex Technology" for Cancer Screening. *Ningen Dock*, 26, pp.911–22.

Okamoto, N., Miyagi, Y., Chiba, A., Akaike, M., Shiozawa, M., Imaizumi, A., Yamamoto, H., Ando, T., Yamakado, M. and Tochikubo, O., 2009. Diagnostic modeling with differences in plasma amino acid profiles between non-cachectic colorectal/breast cancer patients and healthy individuals. *International Journal of Medical Sciences*, 1(1), pp.1–8.

Okamura, K., Takayama, K., Izumi, M., Harada, T., Furuyama, K. and Nakanishi, Y., 2013. Diagnostic value of CEA and CYFRA 21-1 tumor markers in primary lung cancer. *Lung Cancer*, 80(1), pp.45–9.

Opstad, K.S., Bell, B.A., Griffiths, J.R. and Howe, F.A., 2008. An assessment of the effects of sample ischaemia and spinning time on the metabolic profile of brain tumour biopsy specimens as determined by high-resolution magic angle spinning H-1 NMR. *NMR Biomedicine*, 21(10), pp.1138–47.

Opstad, K.S., Bell, B.A., Griffiths, J.R. and Howe, F.A., 2008. An investigation of human brain tumour lipids by high-resolution magic angle spinning 1H MRS and histological analysis. *NMR Biomedicine*, 21, pp.677–85.

Park, Y., Kim, S.B., Wang, B., Blanco, R.A., Le, N.-A., Wu, S., Accardi, C.J., Alexander, R.W., Ziegler, T.R. and Jones, D.P., 2009. Individual variation in macronutrient regulation measured by proton magnetic resonance spectroscopy of human plasma. *American Journal of Physiology*, 297(1), pp.R202–9.

Pesch, B., Kendzia, B., Gustavsson, P., Jöckel, K.-H., Johnen, G., Pohlabeln, H., Olsson, A., Ahrens, W., Gross, I.M., Brüske, I., Wichmann, H.-E., Merletti, F., Richiardi, L., Simonato, L., Fortes, C., Siemiatycki, J., Parent, M.-E., Consonni, D., Landi, M.T., Caporaso, N., Zaridze, D., Cassidy, A., Szeszenia-Dabrowska, N., Rudnai, P., Lissowska, J., Stücker, I., Fabianova, E., Dumitru, R.S., Bencko, V., Foretova, L., Janout, V., Rudin, C.M., Brennan, P., Boffetta, P., Straif, K. and Brüning, T., 2012. Cigarette smoking and lung cancer--relative risk estimates for the major histological types from a pooled analysis of case-control studies. *International Journal of Cancer*, 131(5), pp.1210–9.

Peters, S., Adjei, A.A., Gridelli, C., Reck, M., Kerr, K. and Felip, E., 2012. Metastatic non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 23(Suppl 7), pp.vii56–64.

Pizer, E.S., Wood, F.D., Heine, H.S., Romantsev, F.E., Pasternack, G.R. and Kuhajda, F.P., 1996. Inhibition of fatty acid synthesis delays disease progression in a xenograft model of ovarian cancer. *Cancer Research*, 56(6), pp.1189–93.

Ploessl, K., Wang, L., Lieberman, B.P., Qu, W. and Kung, H.F., 2012. Comparative evaluation of 18F-labeled glutamic acid and glutamine as tumor metabolic imaging agents. *Journal of Nuclear Medicine*, 53(10), pp.1616–24.

Porstmann, T., Griffiths, B., Chung, Y.L., Delpuech, O., Griffiths, J.R., Downward, J. and Schulze, A., 2005. PKB/Akt induces transcription of enzymes involved in cholesterol and fatty acid biosynthesis via activation of SREBP. *Oncogene*, 24(43), pp.6465–81.

Psihogios, N.G., Gazi, I.F., Elisaf, M.S., Seferiadis, K.I. and Bairaktari, E.T., 2008. Gender-related and age-related urinalysis of healthy subjects by NMR-based metabonomics. *NMR Biomedicine*, 21(3), pp.195–207.

Psychogios, N., Hau, D.D., Peng, J., Guo, A.C., Mandal, R., Bouatra, S., Sinelnikov, I., Krishnamurthy, R., Eisner, R., Gautam, B., Young, N., Xia, J., Knox, C., Dong, E., Huang, P., Hollander, Z., Pedersen, T.L., Smith, S.R., Bamforth, F., Greiner, R., McManus, B., Newman, J.W., Goodfriend, T. and Wishart, D.S., 2011. The human serum metabolome. *PLoS ONE*, 6(2), article ID e16957, 23 pages.

Qiu, Y., Cai, G., Su, M., Chen, T., Liu, Y., Xu, Y., Ni, Y., Zhao, A., Cai, S., Xu, L.X. and Jia, W., 2010. Urinary metabonomic study on colorectal cancer. *Journal of Proteome Research*, 9(3), pp.1627–34.

Qiu, Y.P., Cai, G.X., Su, M.M., Chen, T.L., Zheng, X.J., Xu, Y., Ni, Y., Zhao, A.H., Xu, L.X., Cai, S.J. and Jia, W., 2009. Serum metabolite profiling of human colorectal cancer using GC-TOFMS and UPLC-QTOFMS. *Journal Proteome Research*, 8(10), pp.4844–50.

Quint, L.E., Tummala, S., Brisson, L.J., Francis, I.R., Krupnick, A.S., Kazerooni, E.A., Iannettoni, M.D., Whyte, R.I. and Orringer, M.B., 1996. Distribution of distant metastases from newly diagnosed non-small cell lung cancer. *The Annals of Thoracic Surgery*, 62(1), pp.246–50.

Quintás, G., Portillo, N., García-Cañaveras, J.C., Castell, J.V., Ferrer, A. and Lahoz, A., 2011. Chemometric approaches to improve PLSDA model outcome for predicting human non-alcoholic fatty liver disease using UPLC-MS as a metabolic profiling tool. *Metabolomics*, 8(1), pp.86–98.

Raez, L.E., Papadopoulos, K., Ricart, A.D., Chiorean, E.G., Dipaola, R.S., Stein, M.N., Rocha Lima, C.M., Schlesselman, J.J., Tolba, K., Langmuir, V.K., Kroll, S., Jung, D.T., Kurtoglu, M., Rosenblatt, J. and Lampidis, T.J., 2013. A phase I dose-escalation trial of 2-deoxy-D-glucose alone or combined with docetaxel in patients with advanced solid tumors. *Cancer Chemotherapy and Pharmacology*, 71(2), pp.523–30.

Raïs, B., Comin, B., Puigjaner, J., Brandes, J.L., Creppy, E., Saboureau, D., Ennamany, R., Paul Lee, W.-N., Boros, L.G. and Cascante, M., 1999. Oxythiamine and dehydroepiandrosterone induce a G1 phase cycle arrest in Ehrlich's tumor cells through inhibition of the pentose cycle. *FEBS Letters*, 456(1), pp.113–8.

Ramadan, Z., Jacobs, D., Grigorov, M. and Kochhar, S., 2006. Metabolic profiling using principal component analysis, discriminant partial least squares, and genetic algorithms. *Talanta*, 68(5), pp.1683–91.

Rasmussen, L.G., Savorani, F., Larsen, T.M., Dragsted, L.O., Astrup, A. and Engelsen, S.B., 2011. Standardization of factors that influence human urine metabolomics. *Metabolomics*, 7(1), pp.71–83.

Reitzer, L.J., Wice, B.M. and Kennell, D., 1979. Evidence that glutamine, not sugar, is the major energy-source for cultured HeLa cells. *Journal of Biological Chemistry*, 254(8), pp.2669–76.

Reungwetwattana, T., Weroha, S.J. and Molina, J.R., 2012. Oncogenic pathways, molecularly targeted therapies, and highlighted clinical trials in non-small-cell lung cancer (NSCLC). *Clinical Lung Cancer*, 13(4), pp.252–66.

Reynolds, W.F., 2000. *Encyclopedia of Spectroscopy and Spectrometry*, 2nd Edition. Elsevier.

Ridge, C.A., McErlean, A.M. and Ginsberg, M.S., 2013. Epidemiology of Lung Cancer. *Seminars in Interventional Radiology*, 30(2), pp.93–98.

Rivera, M.P. and Mehta, A.C., 2007. Initial diagnosis of lung cancer: ACCP evidence-based clinical practice guidelines (2nd edition). *Chest*, 132(3 Suppl), p.131S–148S.

Rivera, M.P., Mehta, A.C. and Wahidi, M.M., 2013. Establishing the diagnosis of lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest*, 143(Suppl 5), pp.e142S–65S.

Rocha, C., Barros, A., Goodfellow, B., Carreira, I., Gomes, A., Sousa, V., Bernardo, J., Carvalho, L., Gil, A. and Duarte, I., 2014. NMR metabolomics of human lung tumours reveals distinct metabolic signatures for adenocarcinoma and squamous cell carcinoma. *Carcinogenesis*. in press.

Rocha, C.M., Barros, A.S., Gil, A.M., Goodfellow, B.J., Humpfer, E., Spraul, M., Carreira, I.M., Melo, J.B., Bernardo, J., Gomes, A., Sousa, V., Carvalho, L. and Duarte, I.F., 2010. Metabolic profiling of human lung cancer tissue by 1H high resolution magic angle spinning (HRMAS) NMR spectroscopy. *Journal of Proteome Research*, 9(1), pp.319–32.

Rocha, C.M., Carrola, J., Barros, A.S., Gil, A.M., Goodfellow, B.J., Carreira, I.M., Bernardo, J., Gomes, A., Sousa, V., Carvalho, L. and Duarte, I.F., 2011. Metabolic signatures of lung cancer in biofluids: NMR-based metabonomics of blood plasma. *Journal of Proteome Research*, 10(9), pp.4314–24.

Rossi, G., Cavazza, A., Sturm, N., Migaldi, M., Facciolongo, N., Longo, L., Maiorana, A. and Brambilla, E., 2003. Pulmonary carcinomas with pleomorphic, sarcomatoid, or sarcomatous elements: a clinicopathologic and immunohistochemical study of 75 cases. *The American Journal of Surgical Pathology*, 27(3), pp.311–24.

Roux, A., Xu, Y., Heilier, J.-F., Olivier, M.-F., Ezan, E., Tabet, J.-C. and Junot, C., 2012. Annotation of the human adult urinary metabolome and metabolite identification using ultra high performance liquid chromatography coupled to a linear quadrupole ion trap-orbitrap mass spectrometer. *Analytical Chemistry*, 84(15), pp.6429–37.

Rubin, B.P., Skarin, A.T., Pisick, E., Rizk, M. and Salgia, R., 2001. Use of cytokeratins 7 and 20 in determining the origin of metastatic carcinoma of unknown primary, with special emphasis on lung cancer. *European Journal of Cancer Prevention*, 10(1), pp.77–82.

Saccenti, E., Hoefsloot, H.C.J., Smilde, A.K., Westerhuis, J.A. and Hendriks, M.M.W.B., 2013. Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics*, 10(3), pp.361–74.

Salek, R.M., Maguire, M.L., Bentley, E., Rubtsov, D. V, Hough, T., Cheeseman, M., Nunez, D., Sweatman, B.C., Haselden, J.N., Cox, R.D., Connor, S.C. and Griffin, J.L., 2007. A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human. *Physiological Genomics*, 29(2), pp.99–108.

Sanchez-Martinez, C. and Aragon, J.J., 1997. Analysis of phosphofructokinase subunits and isozymes in ascites tumor cells and its original tissue, murine mammary gland. *FEBS Letters*, 409(1), pp.86–90.

Sato, M., Shames, D.S., Gazdar, A.F. and Minna, J.D., 2007. A translational view of the molecular pathogenesis of lung cancer. *Journal of Thoracic Oncology*, 2(4), pp.327–43.

Savorani, F., Tomasi, G. and Engelsen, S.B., 2010. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance*, 202(2), pp.190–202.

Seitz, M., Shukla-Dave, A., Bjartell, A., Touijer, K., Sciarra, A., Bastian, P.J., Stief, C., Hricak, H. and Graser, A., 2009. Functional magnetic resonance imaging in prostate cancer. *European Urology*, 55(4), pp.801–14.

Sekido, Y., Fong, K.M. and Minna, J.D., 2003. Molecular genetics of lung cancer. *Annual Review of Medicine*, 54, pp.73–87.

Semenza, G.L., 2003. Targeting HIF-1 for cancer therapy. *Nature Reviews Cancer*, 3(10), pp.721–32.

Shariff, M.I.F., Gomaa, A.I., Cox, I.J., Patel, M., Williams, H.R.T., Crossey, M.M.E., Thillainayagam, A. V, Thomas, H.C., Waked, I., Khan, S.A. and Taylor-Robinson, S.D., 2011. Urinary metabolic biomarkers of hepatocellular carcinoma in an Egyptian population: a validation study. *Journal of Proteome Research*, 10(4), pp.1828–36.

Shingyoji, M., Iizasa, T., Higashiyama, M., Imamura, F., Saruki, N., Imaizumi, A., Yamamoto, H., Daimon, T., Tochikubo, O., Mitsushima, T., Yamakado, M. and Kimura, H., 2013. The significance and robustness of a plasma free amino acid (PFAA) profile-based multiplex function for detecting lung cancer. *BMC Cancer*, 13(77), 10 pages.

De Silva, S.S., Payne, G.S., Thomas, V., Carter, P.G., Ind, T.E.J. and deSouza, N.M., 2009. Investigation of metabolite changes in the transition from pre-invasive to invasive cervical cancer measured using 1H and 31P magic angle spinning MRS of intact tissue. *NMR Biomedicine*, 22(2), pp.191–8.

Singh, R. and Cuervo, A.M., 2011. Autophagy in the cellular energetic balance. *Cell Metabolism*, 13(5), pp.495–504.

Sitter, B., Bathen, T.F., Tessem, M.-B. and Gribbestad, I.S., 2009. High-resolution magic angle spinning (HR MAS) MR spectroscopy in metabolic characterization of human cancer. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 54(3), pp.239–54.

Sitter, B., Lundgren, S., Bathen, T.F., Halgunset, J., Fjosne, H.E. and Gribbestad, I.S., 2006. Comparison of HR MAS MR spectroscopic profiles of breast cancer tissue with clinical parameters. *NMR Biomedicine*, 19(1), pp.30–40.

Sjøbakk, T.E., Vettukattil, R., Gulati, M., Gulati, S., Lundgren, S., Gribbestad, I.S., Torp, S.H., Bathen, T.F. and Sjobakk, T.E., 2013. Metabolic profiles of brain metastases. *International Journal of Molecular Sciences*, 14(1), pp.2104–18.

Skeaff, C.M., Hodson, L. and McKenzie, J.E., 2006. Dietary-induced changes in fatty acid composition of human plasma, platelet, and erythrocyte lipids follow a similar time course. *The Journal of Nutrition*, 136(3), pp.565–9.

Slupsky, C.M., Cheypesh, A., Chao, D. V, Fu, H., Rankin, K.N., Marrie, T.J. and Lacy, P., 2009. Streptococcus pneumoniae and Staphylococcus aureus p pneumonia induce distinct metabolic responses. *Journal of Proteome Research*, 8(6), pp.3029–36.

Slupsky, C.M., Rankin, K.N., Wagner, J., Fu, H., Chang, D., Weljie, A.M., Saude, E.J., Lix, B., Adamko, D.J., Shah, S., Greiner, R., Sykes, B.D. and Marrie, T.J., 2007. Investigations of the effects of gender, diurnal variation, and age in human urinary metabolomic profiles. *Analytical Chemistry*, 79(18), pp.6995–7004.

Slupsky, C.M., Steed, H., Wells, T.H., Dabbs, K., Schepansky, A., Capstick, V., Faught, W. and Sawyer, M.B., 2010. Urine metabolite analysis offers potential early diagnosis of ovarian and breast cancers. *Clinical Cancer Research*, 16(23), pp.5835–41.

Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. and Siuzdak, G., 2006. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3), pp.779–87.

Sokal, R.R. and Rohlf, F.J., 2012. *Biometry* 4th Edition, W. H. Freeman.

Somashekar, B.S., Kamarajan, P., Danciu, T., Kapila, Y.L., Chinnaiyan, A.M., Rajendiran, T.M. and Ramamoorthy, A., 2011. Magic angle spinning NMR-based metabolic profiling of head and neck squamous cell carcinoma tissues. *Journal of Proteome Research*, 10(11), pp.5232–41.

Spagou, K., Wilson, I.D., Masson, P., Theodoridis, G., Raikos, N., Coen, M., Holmes, E., Lindon, J.C., Plumb, R.S., Nicholson, J.K. and Want, E.J., 2011. HILIC-UPLC-MS for exploratory urinary metabolic profiling in toxicological studies. *Analytical Chemistry*, 83(1), pp.382–90.

Spiro, S.G. and Porter, J.C., 2002. Lung cancer--where are we today? Current advances in staging and nonsurgical treatment. *American Journal of Respiratory and Critical Care Medicine*, 166(9), pp.1166–96.

Sridhar, K.S., Raub, W.A., Duncan, R.C. and Hilsenbeck, S., 1992. The increasing recognition of adenosquamous lung carcinoma (1977-1986). *American Journal of Clinical Oncology*, 15(4), pp.356–62.

Stella, C., Beckwith-Hall, B., Cloarec, O., Holmes, E., Lindon, J.C., Powell, J., van der Ouderaa, F., Bingham, S., Cross, A.J. and Nicholson, J.K., 2006. Susceptibility of human metabolic phenotypes to dietary modulation. *Journal of Proteome Research*, 5(10), pp.2780–8.

Steyerberg, E.W., Bleeker, S.E., Moll, H.A., Grobbee, D.E. and Moons, K.G.M., 2003. Internal and external validation of predictive models: A simulation study of bias and precision in small samples. *Journal of Clinical Epidemiology*, 56(5), pp.441–7.

Stretch, C., Eastman, T., Mandal, R., Eisner, R., Wishart, D.S., Mourtzakis, M., Prado, C.M.M., Damaraju, S., Ball, R.O., Greiner, R. and Baracos, V.E., 2012. Prediction of skeletal muscle and fat mass in patients with advanced cancer using a metabolomic approach. *The Journal of Nutrition*, 142(1), pp.14–21.

Sun, X.-M., Yu, X.-P., Liu, Y., Xu, L. and Di, D.-L., 2012. Combining bootstrap and uninformative variable elimination: Chemometric identification of metabonomic biomarkers by nonparametric analysis of discriminant partial least squares. *Chemometrics and Intelligent Laboratory Systems*, 115, pp.37–43.

Swann, J.R., Spagou, K., Lewis, M., Nicholson, J.K., Glei, D.A., Seeman, T.E., Coe, C.L., Goldman, N., Ryff, C.D., Weinstein, M. and Holmes, E., 2013. Microbial–mammalian cometabolites dominate the age-associated urinary metabolic phenotype in Taiwanese and American populations. *Journal of Proteome Research*, 12(7), pp.3166–80.

Swanson, M.G., Keshari, K.R., Tabatabai, Z.L., Simko, J.P., Shinohara, K., Carroll, P.R., Zektzer, A.S. and Kurhanewicz, J., 2008. Quantification of choline- and ethanolamine-containing metabolites in human prostate tissues using 1H HR-MAS total correlation spectroscopy. *Magnetic Resonance in Medicine*, 60(1), pp.33–40.

Swanson, M.G., Zektzer, A.S., Tabatabai, Z.L., Simko, J., Jarso, S., Keshari, K.R., Schmitt, L., Carroll, P.R., Shinohara, K., Vigneron, D.B. and Kurhanewicz, J., 2006. Quantitative analysis of prostate metabolites using 1H HR-MAS spectroscopy. *Magnetic Resonance in Medicine*, 55(6), pp.1257–64.

Swinnen, J. V, Brusselmans, K. and Verhoeven, G., 2006. Increased lipogenesis in cancer cells: new players, novel targets. *Current Opinion in Clinical Nutrition and Metabolic Care*, 9(4), pp.358–65.

Takahashi, T., Nau, M.M., Chiba, I., Birrer, M.J., Rosenberg, R.K., Vinocour, M., Levitt, M., Pass, H., Gazdar, A.F. and Minna, J.D., 1989. p53: a frequent target for genetic abnormalities in lung cancer. *Science*, 246(4929), pp.491–4.

Takamori, S., Noguchi, M., Morinaga, S., Goya, T., Tsugane, S., Kakegawa, T. and Shimosato, Y., 1991. Clinicopathologic characteristics of adenosquamous carcinoma of the lung. *Cancer*, 67(3), pp.649–54.

Takeda, I., Stretch, C., Barnaby, P., Bhatnager, K., Rankin, K., Fu, H., Weljie, A., Jha, N. and Slupsky, C., 2009. Understanding the human salivary metabolome. *NMR Biomedicine*, 22(6), pp.577–84.

Tautenhahn, R., Böttcher, C. and Neumann, S., 2008. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics*, 9(1), article ID 504, 16 pages.

Taylor, G., 1964. Disintegration of water drops in an electric field. *Proceedings of the Royal Society A*, 280(1382), pp.383–97.

Teahan, O., Gamble, S., Holmes, E., Waxman, J., Nicholson, J.K., Bevan, C. and Keun, H.C., 2006. Impact of analytical bias in metabonomic studies of human blood serum and plasma. *Analytical Chemistry*, 78(13), pp.4307–18.

Terasaki, H., Niki, T., Matsuno, Y., Yamada, T., Maeshima, A., Asamura, H., Hayabuchi, N. and Hirohashi, S., 2003. Lung adenocarcinoma with mixed bronchioloalveolar and invasive components: clinicopathological features, subclassification by extent of invasive foci, and immunohistochemical characterization. *The American Journal of Surgical Pathology*, 27(7), pp.937–51.

Tessem, M.-B., Swanson, M.G., Keshari, K.R., Albers, M.J., Joun, D., Tabatabai, Z.L., Simko, J.P., Shinohara, K., Nelson, S.J., Vigneron, D.B., Gribbestad, I.S. and Kurhanewicz, J., 2008. Evaluation of lactate and alanine as metabolic biomarkers of prostate cancer using 1H HR-MAS spectroscopy of biopsy tissues. *Magnetic Resonance Medicine*, 60(3), pp.510–6.

Thallion Pharmaceuticals, 2009. *Drug development TLN-232*, Quebec, Canada.

Thun, M.J., Hannan, L.M., Adams-Campbell, L.L., Boffetta, P., Buring, J.E., Feskanich, D., Flanders, W.D., Jee, S.H., Katanoda, K., Kolonel, L.N., Lee, I-M., Marugame, T.,Palmer, J.R., Riboli, E., Sobue, T., Avila-Tang, E., Wilkens, L.R., Samet, J.M., 2008. Lung cancer occurrence in never-smokers: an analysis of 13 cohorts and 22 cancer registry studies. *PLoS Medicine*, 5 (9), article ID e185, 14 pages.

Tiziani, S., Lopes, V. and Günther, U.L., 2009. Early stage diagnosis of oral cancer using 1H NMR–based metabolomics. *Neoplasia*, 11(3), pp.269–76.

Tolun, A.A., Zhang, H., Il'yasova, D., Sztáray, J., Young, S.P. and Millington, D.S., 2010. Allantoin in human urine quantified by ultra-performance liquid chromatography-tandem mass spectrometry. *Analytical Biochemistry*, 402(2), pp.191–3.

Tomashefski, J.F., Connors, A.F., Rosenthal, E.S. and Hsiue, I.L., 1990. Peripheral vs central squamous cell carcinoma of the lung. A comparison of clinical features, histopathology, and survival. *Archives of Pathology & Laboratory Medicine*, 114(5), pp.468–74.

Tong, X., Zhao, F. and Thompson, C.B., 2009. The molecular determinants of de novo nucleotide biosynthesis in cancer cells. *Current Opinion in Genetics & Development*, 19(1), pp.32–7.

Tran, T.T., Nguyen, T.M.P., Nguyen, B.N. and Phan, V.C., 2008. Changes of serum glycoproteins in lung cancer patients. *Journal of Proteomics and Bioinformatics*, 1, pp.11–6.

Traverso, N., Ricciarelli, R., Nitti, M., Marengo, B., Furfaro, A.L., Pronzato, M.A., Marinari, U.M. and Domenicotti, C., 2013. Role of glutathione in cancer progression and chemoresistance. *Oxidative Medicine and Cellular Longevity*, 2013, article ID 972913, 10 pages.

Travis, W.D., 2010a. Sarcomatoid neoplasms of the lung and pleura. *Archives of Pathology & Laboratory Medicine*, 134(11), pp.1645–58.

Travis, W.D., 2010b. Advances in neuroendocrine lung tumors. *Annals of Oncology*, 21(Suppl 7), pp.vii65–71.

Travis, W.D., 2011. Pathology of Lung Cancer. *Clinics in Chest Medicine*, 32(4), pp.669–92.

Travis, W.D., Brambilla, E., Müller-Hermelink, H.K. and Harris, C., 2004. *World Health Organization classification of tumours; tumours of lung, pleura, thymus and heart*, Lyon, France: IARC Press.

Travis, W.D., Brambilla, E. and Riely, G.J., 2013. New pathologic classification of lung cancer: relevance for clinical practice and clinical trials. *Journal of Clinical Oncology*, 31(8), pp.992–1001.

Travis, W.D., Linnoila, R.I., Tsokos, M.G., Hitchcock, C.L., Cutler, G.B., Nieman, L., Chrousos, G., Pass, H. and Doppman, J., 1991. Neuroendocrine tumors of the lung with proposed criteria for large-cell neuroendocrine carcinoma. An ultrastructural, immunohistochemical, and flow cytometric study of 35 cases. *The American Journal of Surgical Pathology*, 15(6), pp.529–53.

Travis, W.D., Rush, W., Flieder, D.B., Falk, R., Fleming, M. V, Gal, A.A. and Koss, M.N., 1998. Survival analysis of 200 pulmonary neuroendocrine tumors with clarification of criteria for atypical carcinoid and its separation from typical carcinoid. *The American Journal of Surgical Pathology*, 22(8), pp.934–44.

Travis, W.D., Travis, L.B. and Devesa, S.S., 1995. Lung cancer. *Cancer*, 75(Suppl 1), pp.191–202.

Trump, S., Laudi, S., Unruh, N., Goelz, R. and Leibfritz, D., 2006. 1H-NMR metabolic profiling of human neonatal urine. *Magnetic Resonance Materials in Physics*, 19, pp.305–12.

Trygg, J., Holmes, E. and Lundstedt, T., 2007. Chemometrics in Metabonomics. *Journal of Proteome Research*, 6(2), pp.469–79.

Trygg, J. and Wold, S., 2002. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, 16(3), pp.119–28.

Tugnoli, V., Mucci, A., Schenetti, L., Righi, V., Calabrese, C., Fabbri, A., Di Febo, G. and Tosi, M.R., 2006. Ex vivo HR-MAS magnetic resonance spectroscopy of human gastric adenocarcinomas: a comparison with healthy gastric mucosa. *Oncology Reports*, 16(3), pp.543–53.

Ulivi, P., Mercatali, L., Zoli, W., Dell'amore, D., Poletti, V., Casoni, G.L., Scarpi, E., Flamini, E., Amadori, D. and Silvestrini, R., 2008. Serum free DNA and COX-2 mRNA expression in peripheral blood for lung cancer detection. *Thorax*, 63(9), pp.843–4.

Ulrich, E.L., Akutsu, H., Doreleijers, J.F., Harano, Y., Ioannidis, Y.E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., Nakatani, E., Schulte, C.F., Tolmie, D.E., Kent Wenger, R., Yao, H. and Markley, J.L., 2008. BioMagResBank. *Nucleic Acids Research*, 36(Database issue), pp.D402–8.

Urayama, S., Zou, W., Brooks, K. and Tolstikov, V., 2010. Comprehensive mass spectrometry based metabolic profiling of blood plasma reveals potent

discriminatory classifiers of pancreatic cancer. *Rapid Communications in Mass Spectrometry*, 24(5), pp.613–20.

Van, Q., Veenstra, T. and Issaq, H., 2011. Metabolic profiling for the detection of bladder cancer. *Current Urology Reports*, 12(1), pp.34–40.

Vansteenkiste, J., De Ruysscher, D., Eberhardt, W.E.E., Lim, E., Senan, S., Felip, E. and Peters, S., 2013. Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 24(Suppl 6), pp.vi89–98.

Vaughan, A.A., Dunn, W.B., Allwood, J.W., Wedge, D.C., Blackhall, F.H., Whetton, A.D., Dive, C. and Goodacre, R., 2012. Liquid chromatography-mass spectrometry calibration transfer and metabolomics data fusion. *Analytical Chemistry*, 84(22), pp.9848–57.

Vaupel, P., Kallinowski, F. and Okunieff, P., 1989. Blood flow, oxygen and nutrient supply, and metabolic microenvironment of human tumors: A review. *Cancer Research*, 49(23), pp.6449–65.

Veering, B.T., Burm, A.G., Souverijn, J.H., Serree, J.M. and Spierdijk, J., 1990. The effect of age on serum concentrations of albumin and alpha 1-acid glycoprotein. *British Journal of Clinical Pharmacology*, 29(2), pp.201–6.

Veselkov, K.A., Lindon, J.C., Ebbels, T.M.D., Crockford, D., Volynkin, V. V, Holmes, E., Davies, D.B. and Nicholson, J.K., 2009. Recursive segment-wise peak alignment of biological 1H NMR spectra for improved metabolic biomarker recovery. *Analytical Chemistry*, 81(1), pp.56–66.

Veselkov, K.A., Vingara, L.K., Masson, P., Robinette, S.L., Want, E., Li, J. V, Barton, R.H., Boursier-Neyret, C., Walther, B., Ebbels, T.M., Pelczer, I., Holmes, E., Lindon, J.C. and Nicholson, J.K., 2011. Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Analytical Chemistry*, 83(15), pp.5864–72.

Vizan, P., Mazurek, S. and Cascante, M., 2008. Robust metabolic adaptation underlying tumor progression. *Metabolomics*, 4(1), pp.1–12.

Vu, T.N., Valkenborg, D., Smets, K., Verwaest, K.A., Dommisse, R., Lemière, F., Verschoren, A., Goethals, B. and Laukens, K., 2011. An integrated workflow for

robust alignment and simplified quantitative analysis of NMR spectrometry data. *BMC Bionformatics*, 12, article ID 405, 14 pages.

Walsh, M.C., Brennan, L., Malthouse, J.P.G., Roche, H.M. and Gibney, M.J., 2006. Effect of acute dietary standardization on the urinary, plasma, and salivary metabolomic profiles of healthy humans. *The American Journal of Clinical Nutrition*, 84(3), pp.531–9.

Wang, R., Wang, G., Zhang, N., Li, X. and Liu, Y., 2013. Clinical evaluation and cost-effectiveness analysis of serum tumor markers in lung cancer. *BioMed Research International*, 2013, article ID 195692, 7 pages.

Wang, X., Christiani, D.C., Wiencke, J.K., Fischbein, M., Xu, X., Cheng, T.J., Mark, E., Wain, J.C. and Kelsey, K.T., 1995. Mutations in the p53 gene in lung cancer are associated with cigarette smoking and asbestos exposure. *Cancer Epidemiology, Biomarkers & Prevention*, 4(5), pp.543–8.

Wang, Y., Cloarec, O., Tang, H., Lindon, J.C., Holmes, E., Kochhar, S. and Nicholson, J.K., 2008. Magic Angle Spinning NMR and 1H-31P Heteronuclear Statistical Total Correlation Spectroscopy of Intact Human Gut Biopsies. *Analytical Chemistry*, 80(4), pp.1058–66.

Wang, Y. and Griffiths, W.J., 2008. Mass spectrometry for metabolite identification. In W. J. Griffiths, 1st Edition. *Metabolomics, Metabonomics and Metabolite Profiling*. RSC Publishing.

Want, E.J., Nordström, A., Morita, H. and Siuzdak, G., 2007. From exogenous to endogenous: the inevitable imprint of mass spectrometry in metabolomics. *Journal of Proteome Research*, 6(2), pp.459–68.

Want, E.J., Wilson, I.D., Gika, H., Theodoridis, G., Plumb, R.S., Shockcor, J., Holme, E. and Nicholson, J.K., 2010. Global metabolic profiling procedures for urine using UPLC–MS. *Nature Protocols*, 5(6), pp.1005–18.

Warburg, O., 1956. On origin of cancer cells. *Science*, 123(3191), pp.309–14.

Warburg, O., Posener, K. and Negelein, E., 1924. Uber den Stoffwechsel der Karzinomzellen. *Biochemische Zeitschrift*, 152, pp.309–44.

Weinberg, R., 2013. *The biology of cancer*. New York, USA: Garland Science.

Wen, T., Gao, L., Wen, Z., Wu, C., Tan, C.S., Toh, W.Z. and Ong, C.N., 2013. Exploratory investigation of plasma metabolomics in human lung adenocarcinoma. *Molecular bioSystems*, 9(9), pp.2370–8.

Westerhuis, J.A., Hoefsloot, H.C.J., Smit, S., Vis, D.J., Smilde, A.K., Velzen, E.J.J., Duijnhoven, J.P.M. and Dorsten, F.A., 2008. Assessment of PLSDA cross validation. *Metabolomics*, 4(1), pp.81–9.

Wiklund, S., Johansson, E., Sjostrom, L., Mellerowicz, E.J., Edlund, U., Shockcor, J.P., Gottfries, J., Moritz, T. and Trygg, J., 2008. Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. *Analytical Chemistry*, 80(1), pp.115–22.

Wilm, M., 2011. Principles of Electrospray Ionization. *Molecular & Cellular Proteomics*, 10(7), 8 pages.

Winnike, J.H., Busby, M.G., Watkins, P.B. and O'Connell, T.M., 2009. Effects of a prolonged standardized diet on normalizing the human metabolome. *The American Journal of Clinical Nutrition*, 90(6), pp.1496–501.

Wise, D.R., DeBerardinis, R.J., Mancuso, A., Sayed, N., Zhang, X.-Y., Pfeiffer, H.K., Nissim, I., Daikhin, E., Yudkoff, M., McMahon, S.B. and Thompson, C.B., 2008. Myc regulates a transcriptional program that stimulates mitochondrial glutaminolysis and leads to glutamine addiction. *Proceedings of the National Academy of Sciences*, 105(48), pp.18782–7.

Wishart, D.S., 2013. Exploring the human metabolome by Nuclear Magnetic Resonance and Mass Spectrometry. In N. W. Lutz, J. V. Sweedler, & R. A. Wevers, eds. *Methodologies for Metabolomics*. Cambridge University Press, pp. 3–29.

Wishart, D.S., Tzur, D., Knox, C., Eisner, R., Guo, A.C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M.-A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D.D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G.E., MacInnis, G.D., Weljie, A.M., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B.D., Vogel, H.J. and Querengesser, L., 2007. HMDB: the human metabolome database. *Nucleic Acids Research*, 35(Database issue), pp.D521-D526.

Wold, S., Sjöström, M. and Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2), pp.109–30.

Wood, T., 1986. Physiological functions of the pentose-phosphate pathway. *Cell Biochemistry and Function*, 4(4), pp.241–7.

Wu, M., Xu, Y., Fitch, W.L., Zheng, M., Merritt, R.E., Shrager, J.B., Zhang, W., Dill, D.L., Peltz, G. and Hoang, C.D., 2013. Liquid chromatography/mass spectrometry methods for measuring dipeptide abundance in non-small-cell lung cancer. *Rapid Communications in Mass Spectrometry*, 27(18), pp.2091–8.

Wu, Q., Wang, Y., Gu, X., Zhou, J., Zhang, H., Lv, W., Chen, Z. and Yan, C., 2014. Urinary metabolomic study of non-small-cell lung carcinoma based on ultra high performance liquid chromatography coupled with quadrupole time-of-flight mass spectrometry. *Journal of Separation Science*, 37(14), pp.1728–35.

Wyss, M. and Kaddurah-Daouk, R., 2000. Creatine and creatinine metabolism. *Physiological Reviews*, 80(3), pp.1108–213.

Xiang, D., Zhang, B., Doll, D., Shen, K., Kloecker, G. and Freter, C., 2013. Lung cancer screening: from imaging to biomarker. *Biomarker Research*, 1(1), article ID 4, 9 pages.

Xiao, C., Hao, F., Qin, X., Wang, Y. and Tang, H., 2009. An optimized buffer system for NMR-based urinary metabonomics with effective pH control, chemical shift consistency and dilution minimization. *Analyst*, 134, pp.916–925.

Xie, G.X., Qiu, Y.P., Shi, P., Zheng, X.J., Su, M.M., Zhao, A.H., Zhou, Z.T., Jia, W. and Chen, T., 2012. Urine metabolite profiling offers potential early diagnosis of oral cancer. *Metabolomics*, 8, pp.220–31.

Xie, H., Valera, V.A., Merino, M.J., Amato, A.M., Signoretti, S., Linehan, W.M., Sukhatme, V.P. and Seth, P., 2009. LDH-A inhibition, a therapeutic strategy for treatment of hereditary leiomyomatosis and renal cell cancer. *Molecular Cancer Therapeutics*, 8(3), pp.626–35.

Xu, Q.S. and Liang, Y.Z., 2001. Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, 56(1), pp.1–11.

Xu, T., Holzapfel, C., Dong, X., Bader, E., Yu, Z., Prehn, C., Perstorfer, K., Jaremek, M., Roemisch-Margl, W., Rathmann, W., Li, Y., Wichmann, H.E., Wallaschofski, H., Ladwig, K.H., Theis, F., Suhre, K., Adamski, J., Illig, T., Peters, A. and Wang-

Sattler, R., 2013. Effects of smoking and smoking cessation on human serum metabolite profile: results from the KORA cohort study. *BMC Medicine*, 11(1), article ID 60, 14 pages.

Yang, C., Sudderth, J., Dang, T., Bachoo, R.M., Bachoo, R.G., McDonald, J.G. and DeBerardinis, R.J., 2009. Glioblastoma cells require glutamate dehydrogenase to survive impairments of glucose metabolism or Akt signaling. *Cancer Research*, 69(20), pp.7986–93.

Yang, Q., Shi, X., Wang, Y., Wang, W., He, H., Lu, X. and Xu, G., 2010. Urinary metabonomic study of lung cancer by a fully automatic hyphenated hydrophilic interaction/RPLC-MS system. *Journal of Separation Science*, 33(10), pp.1495–503.

Yang, Y., Li, C., Nie, X., Feng, X., Chen, W., Yue, Y., Tang, H. and Deng, F., 2007. Metabonomic studies of human hepatocellular carcinoma using high-resolution magic-angle spinning 1H NMR spectroscopy in conjunction with multivariate data analysis. *Journal of Proteome Research*, 6(7), pp.2605–14.

Yang, Y., Wang, L., Wang, S., Liang, S., Chen, A., Tang, H., Chen, L. and Deng, F., 2013. Study of metabonomic profiles of human esophageal carcinoma by use of high-resolution magic-angle spinning 1H NMR spectroscopy and multivariate data analysis. *Analytical and Bioanalytical Chemistry*, 405(10), pp.3381–9.

Yao, X., Gomes, M.M., Tsao, M.S., Allen, C.J., Geddie, W. and Sekhon, H., 2012. Fine-needle aspiration biopsy versus core-needle biopsy in diagnosing lung cancer: a systematic review. *Current oncology*, 19(1), pp.e16–27.

Yatabe, Y., Mitsudomi, T. and Takahashi, T., 2002. TTF-1 expression in pulmonary adenocarcinomas. *The American Journal of Surgical Pathology*, 26(6), pp.767–73.

Ye, J., Mancuso, A., Tong, X., Ward, P.S., Fan, J., Rabinowitz, J.D. and Thompson, C.B., 2012. Pyruvate kinase M2 promotes de novo serine synthesis to sustain mTORC1 activity and cell proliferation. *Proceedings of the National Academy of Sciences of the United States of America*, 109(18), pp.6904–9.

Yeo, E.-J., Chun, Y.-S., Cho, Y.-S., Kim, J., Lee, J.-C., Kim, M.-S. and Park, J.-W., 2003. YC-1: A Potential Anticancer Drug Targeting Hypoxia-Inducible Factor 1. *JNCI Journal of the National Cancer Institute*, 95(7), pp.516–525.

Yeung, S.J., Pan, J. and Lee, M.-H., 2008. Roles of p53, MYC and HIF-1 in regulating glycolysis - the seventh hallmark of cancer. *Cellular and Molecular Life Sciences*, 65(24), pp.3981–99.

Yokota, H., Guo, J., Matoba, M., Higashi, K., Tonami, H. and Nagao, Y., 2007. Lactate, choline and creatine levels measured by vitro 1H-MRS as prognostic parameters in patients with non-small-cell lung cancer. *Journal of Magnetic Resonance Imaging*, 25(5), pp.992–9.

Young, C.D. and Anderson, S.M., 2008. Sugar and fat - that's where it's at: metabolic changes in tumors. *Breast Cancer Research*, 10(1), article ID 202, 9 pages.

Yu, J.B., Decker, R.H., Detterbeck, F.C. and Wilson, L.D., 2010. Surveillance epidemiology and end results evaluation of the role of surgery for stage I small cell lung cancer. *Journal of Thoracic Oncology*, 5(2), pp.215–9.

Yu, Z., Zhai, G., Singmann, P., He, Y., Xu, T., Prehn, C., Römisch-Margl, W., Lattka, E., Gieger, C., Soranzo, N., Heinrich, J., Standl, M., Thiering, E., Mittelstraß, K., Wichmann, H.-E., Peters, A., Suhre, K., Li, Y., Adamski, J., Spector, T.D., Illig, T. and Wang-Sattler, R., 2012. Human serum metabolic profiles are age dependent. *Aging Cell*, 11(6), pp.960–7.

Zeng, X., Hood, B.L., Sun, M., Conrads, T.P., Day, R.S., Weissfeld, J.L., Siegfried, J.M. and Bigbee, W.L., 2010. Lung cancer serum biomarker discovery using glycoprotein capture and liquid chromatography mass spectrometry. *Journal of Proteome Research*, 9(12), pp.6440–9.

Zerbe, O. and Jurt, S., 2014. *Applied NMR spectroscopy for chemists and life scientists*, Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA.

Zhang, J.A., Liu, L.Y., Wei, S.W., Gowda, G.A.N., Hammoud, Z., Kesler, K.A. and Raftery, D., 2011. Metabolomics study of esophageal adenocarcinoma. *Journal of Thoracic and Cardiovascular Surgery*, 141(2), pp.469–U1048.

Zhang, L., Jin, H., Guo, X., Yang, Z., Zhao, L., Tang, S., Mo, P., Wu, K., Nie, Y., Pan, Y. and Fan, D., 2012. Distinguishing pancreatic cancer from chronic pancreatitis and healthy individuals by H-1 nuclear magnetic resonance-based metabonomic profiles. *Clinical Biochemistry*, 45(13-14), pp.1064–9.

Zhang, S., Liu, L., Steffen, D., Ye, T. and Raftery, D., 2011. Metabolic profiling of gender: Headspace-SPME/GC–MS and 1H NMR analysis of urine. *Metabolomics*, 8(2), pp.1–12.

Zhang, Z.Y., Qiu, Y.P., Hua, Y.Q., Wang, Y.H., Chen, T.L., Zhao, A.H., Chi, Y., Pan, L., Hu, S., Li, J.A., Yang, C.W., Li, G.D., Sun, W., Cai, Z.D. and Jia, W., 2010. Serum and urinary metabonomic study of human osteosarcoma. *Journal of Proteome Research*, 9(9), pp.4861–8.

Zhou, X.-M., He, C.-C., Liu, Y.-M., Zhao, Y., Zhao, D., Du, Y., Zheng, W.-Y. and Li, J.-X., 2012. Metabonomic classification and detection of small molecule biomarkers of malignant pleural effusions. *Analytical and Bioanalytical Chemistry*, 404(10), pp.3123–33.

Zhu, A., Lee, D. and Shim, H., 2011. Metabolic positron emission tomography imaging in cancer detection and therapy response. *Seminars in Oncology*, 38(1), pp.55–69.

Zira, A.N., Theocharis, S.E., Mitropoulos, D., Migdalis, V. and Mikros, E., 2010. 1H NMR metabonomic analysis in renal cell carcinoma: a possible diagnostic tool. *Journal of Proteome Research*, 9(8), pp.4038–44.

Zou, K.H., Tuncali, K. and Silverman, S.G., 2003. Correlation and simple linear regression. *Radiology*, 227(3), pp.617–22.

ANNEXES

ANNEX I: Histological classification of lung tumours

Table A1 World Health Organisation (WHO) histological classification of malignant epithelial tumours of the lung (Travis et al. 2004).

Histological classification of tumours
Squamous cell carcinoma
Papillary
Clear cell
Small cell
Basaloid
Small cell carcinoma
Combined small cell carcinoma
Adenocarcinoma
Adenocarcinoma, mixed subtype
Acinar adenocarcinoma
Papillary adenocarcinoma
Bronchioloalveolar carcinoma
Nonmucinous
Mucinous
Mixed nonmucinous and mucinous or indeterminate
Solid adenocarcinoma with mucin production
Fetal adenocarcinoma
Mucinous ('colloid') carcinoma
Mucinous cystadenocarcinoma
Signet ring adenocarcinoma
Clear cell adenocarcinoma
Large cell carcinoma
Large cell neuroendocrine carcinoma
Combined large cell neuroendocrine carcinoma
Basaloid carcinoma
Lymphoepithelioma-lyke carcinoma
Clear cell carcinoma
Large cell carcinoma with rhabdoid phenotype
Adenosquamous carcinoma
Sarcomatoid carcinoma
Pleomorphic carcinoma
Spindle cell carcinoma
Giant cell carcinoma
Carcinosarcoma
Pulmonary blastoma
Carcinoid tumour
Typical carcinoid
Atypical carcinoid

ANNEX II: TNM classification and stage grouping of lung tumours

Table A2 TNM (Tumour, Node, Metastasis) classification of carcinomas of the lung (Travis et al. 2004).

TNM classification of the lung	
T	**Primary tumour**
Tx	Primary tumour cannot be assessed, or tumour assessed by positive cytology only
T0	No evidence of primary tumour
Tis	Carcinoma in situ
T1	≤3 cm
T1a	≤2 cm
T1b	>2-3 cm
T2	Main bronchus ≥2 cm from carina invades pleura, partial atelectasis
T2a	>3-5 cm
T2b	>5-7 cm
T3	>7 cm; chest wall, diaphragm, pericardium, mediastinal pleura, main bronchus >2 cm from carina, total atelectasis, separate nodule(s) in the same lobe
T4	Mediastinum, heart, great vessels, carina, trachea, oesophagus, vertebra; separate tumour nodule(s) in a ipsilateral lobe
N	**Regional lymph nodes**
Nx	Regional lymph nodes cannot be assessed
N0	No regional lymph node metastasis
N1	Ipsilateral peribronchial, ipsilateral hilar
N2	Subcarinal, ipsilateral mediastinal
N3	Contralateral mediastinal or hilar, scalene, or supraclavicular
M	**Distant metastasis**
Mx	Distant metastasis cannot be assessed
M0	No distant metastasis
M1	Presence of distant metastasis
M1a	Separate tumour nodule(s) in a contralateral lobe; pleural nodules or malignant pleural, or pericardial effusion
M1b	Distant metastasis

Table A3 Stage grouping of carcinomas of the lung (Travis et al. 2004).

Stage grouping			
Occult carcinoma	Tx	N0	M0
Stage 0	Tis	N0	M0
Stage IA	T1a,b	N0	M0
Stage IB	T2a	N0	M0
Stage IIA	T2b	N0	M0
	T1a,b	N1	M0
	T2a	N1	M0
Stage IIB	T2b	N1	M0
	T3	N0	M0
Stage IIIA	T1a,b, T2a,b	N2	M0
	T3	N1, N2	M0
	T4	N0, N1	M0
Stage IIIB	T4	N2	M0
	Any T	N3	M0
Stage IV	Any T	Any N	M1

ANNEX III: List of lung cancer metabolic profiling studies available in the literature

Table A4 Main characteristics of the metabolic profiling studies of lung cancer (LC) reported in the literature. Only studies involving the analysis of human tissues and/or biofluids are included (studies using animal models or cultured cells are not contemplated in this list).

Subject groups (no. samples)			Analytical technique	MVA methods	Metabolic profiles compared				Reference
controls	LC patients	other			cancer vs. control	histological types	disease stages	other	
Tissue extracts									
27	30		^{1}H NMR	Discriminant factor analysis	✓	✓			Hanaoka et al. 1993
0	19		^{1}H NMR	n.a.				^{18}F-FDG uptake; survival probability	Guo et al. 2004
20	21		^{1}H NMR	n.a.	✓			recurrence; survival probability	Yokota et al. 2007
10	10		1D and 2D NMR; GC-MS	n.a.	✓				Fan et al. 2009
9	9	PC	CE-ToF MS	PCA	✓	✓		PC	Kami et al. 2013
Tissue									
7[a]	14		^{1}H HRMAS NMR	PCA, CA	✓	✓			Jordan et al. 2010
12	12		^{1}H HRMAS NMR	PCA, HCA	✓				Rocha et al. 2010
24	24		^{1}H HRMAS NMR	PCA, PLS-DA	✓	✓			Duarte et al. 2010
0	17		^{1}H HRMAS NMR	PCA, OPLS-DA		✓	✓	tumour location	Chen et al. 2011
56	56		^{1}H HRMAS NMR	PCA, PLS-DA, PLS-R	✓	✓	✓	tumour and necrotic fractions	Rocha et al. 2014
Blood serum and plasma									
423	141		HPLC-MS	PCA, MLR	✓	✓	✓		Maeda et al. 2010
200	996	GCA, CRC, BC, PC	HPLC-MS	LDA	✓		✓		Miyagi et al. 2011

Table A4 (continued)

Subject groups (no. samples)			Analytical technique	MVA methods	Metabolic profiles compared				Reference
controls	LC patients	other			cancer vs. control	histological types	disease stages	other	
Blood serum and plasma									
3849	171		HPLC-MS	Several discrimiant functions	✓				Shingyoji et al. 2013
12	12		UPLC-Q-ToF MS	PCA	✓				Dong et al. 2010
495	58		DI-FTICR MS	PLS-DA	✓		✓	Gender	Guo et al. 2012
7	14		^{1}H HRMAS NMR	PCA, CA	✓	✓			Jordan et al. 2010
78	85		^{1}H NMR	PCA, PLS-DA, OPLS-DA	✓	✓	✓	Gender, age, smoking habits	Rocha et al. 2011
29	33		GC-MS	PLS-DA	✓	✓	✓		Hori et al. 2011
28	66		UPLC-MS	PLS-DA	✓			Radiotherapy periods	Cai et al. 2011
100	100		DI-MS	PCA, SVM	✓		✓		Lokhov et al. 2012
0	29		UPLC-MS	PCA, PLS-R				Survival time	Vaughan et al. 2012
28	31		GC-MS, LC-MS	OPLS-DA	✓				Wen et al. 2013
Urine									
	2	RC, HNC	HPLC, LC-IC-MS, FTICR-MS	n.a.					Bullinger et al. 2008
22	19		RRLC-APCI/APPI-MS	PLS-DA, OPLS-DA	✓				An et al. 2010
35	32		HILIC/RPLC-MS	OSC PLS-DA	✓				Yang et al. 2010
20	20		UHPL-MS	PCA, OPLS-DA	✓				Wu et al. 2014
536	469		UPLC-MS	Unconditional logistic regression	✓		✓	Smoking status	Mathe et al. 2014
54	71		^{1}H NMR	PCA, PLS-DA, OPLS-DA	✓			Gender, age, smoking habits	Carrola et al. 2011

Table A4 (continued)

Subject groups (no. samples)			Analytical technique	MVA methods	Metabolic profiles compared				Reference
controls	LC patients	other			cancer vs. control	histological types	disease stages	other	
Urine									
0	93 (LC+CC)		^{1}H NMR	PLS-DA, decision tree, SVM, PIA and other				loss of muscle mass	Eisner et al. 2011
0	55 (LC+CC)		^{1}H NMR, DI-MS	PLS-DA, SVM, LASSO				loss of muscle mass	Stretch et al. 2012
Other (EBC and/or BALF)									
22	17		EIA	n.a.	✓			Smoking habits	Ciebiada et al. 2012
40	41		^{1}H NMR	PCA, PLS-DA, OPL-DA, DFA	✓				Zhou et al. 2012
20	20		HS-SPME-GC-MS	n.a.	✓				Liu et al. 2014

[a] Serum control samples; n.a. not applicable; BALF: bronchoalveolar lavage fluid; BC: breast cancer; CA: canonical analysis; CC: colon cancer; CE: capillary electrophoresis; CRC: colorectal cancer; DFA: discriminant function analysis; DI: direct injection; EBC: exhaled breath condensate; EIA: enzyme immunoassay; FDG: fluorodeoxyglucose; FTICR: Fourier transform ion cyclotron resonance; GC: gas chromatography; GCA: gastric cancer; HCA: hierarchical cluster analysis; HNC: head and neck cancer; HPLC: high performance liquid chromatography; HS-SPME: headspace – solid phase microextraction; LASSO: least absolute shrinkage and selection operator, LC: lung cancer; MLR: multiple logistic regression; MS: mass spectrometry; MVA: multivariate analysis; NMR: nuclear magnetic resonance; OPLS-DA: orthogonal projection to latent structures –discriminant analysis; PC: prostate cancer; PCA: principal component analysis; PLS-DA: partial least squares – discriminant analysis; PLS-R: partial least squares regression; RC: rectal cancer; SVM: support vector machines; ToF: time of flight; UHPLC or UPLC: ultrahigh performance liquid chromatography.

ANNEX IV: Demographic and histological information on lung cancer patients enrolled in this study

Table A5 Demographic and histological information of lung cancer patients enrolled in this work.

Patient no.	Diagnosis	Stage	Age	Gender	Patient no.	Diagnosis	Stage	Age	Gender	Patient no.	Diagnosis	Stage	Age	Gender
1	Adenocarcinoma	IIA	62	M	24	Adenocarcinoma	IB	55	F	47	Epidermoid	IIB	69	M
2	Adenocarcinoma	n.a.	74	F	25	Adenocarcinoma	IA	42	F	48	Adenocarcinoma	IIB	51	F
3	Adenocarcinoma	IA	68	F	26	Adenocarcinoma	IIIB	69	F	49	Carcinoid	IIA	62	M
4	Carcinoid	IA	66	M	27	Adenocarcinoma	IA	75	M	50	Adenocarcinoma	IB	55	F
5	Adenocarcinoma	IB	57	F	28	Adenocarcinoma	IB	64	F	51	Sarcomatoid	IB	73	M
6	Adenocarcinoma	IIA	51	M	29	Epidermoid	IIB	66	M	52	Adenocarcinoma	IB	73	M
7	Adenocarcinoma	IB	75	F	30	Sarcomatoid	inconclusive	68	M	53	Adenocarcinoma	IIA	61	M
8	Adenocarcinoma	IIB	71	M	31	LCC	IA	61	M	54	Epidermoid	IA	43	M
9	Adenocarcinoma	inconclusive	80	F	32	Carcinoid	IIB	61	F	55	SCC	IIB	52	F
10	Adenocarcinoma	IIB	49	M	33	Adenocarcinoma	IB	49	M	56	Epidermoid	IIB	66	M
11	Epidermoid	IA	59	M	34	Epidermoid	IB	65	F	57	Adenocarcinoma	IIB	70	F
12	Adenocarcinoma	IB	81	M	35	Epidermoid	IB	58	M	58	Carcinoid	IB	57	F
13	Adenocarcinoma	IIB	59	M	36	Adenocarcinoma	IB	71	M	59	Adenocarcinoma	IA	54	F
14	Sarcomatoid	IA	58	F	37	Adenocarcinoma	IA	68	M	60	Carcinoid	IA	58	F
15	LCC	IIIA	79	M	38	Adenocarcinoma	IIIA	68	M	61	Epidermoid	IB	71	M
16	Epidermoid	IB	58	M	39	Sarcomatoid	IIB	46	M	62	Epidermoid	IA	66	M
17	Sarcomatoid	IIIA	63	M	40	Epidermoid	IA	63	M	63	SCC	IIIA	63	F
18	Epidermoid	IB	74	M	41	Adenocarcinoma	n.a.	75	M	64	Adenocarcinoma	IA	71	F
19	Epidermoid	IIA	62	M	42	Sarcomatoid	IA	57	M	65	Adenocarcinoma	IIB	64	F
20	Adenocarcinoma	IIB	77	F	43	Epidermoid	IIB	65	M	66	Sarcomatoid	IA	61	M
21	Carcinoid	IA	71	M	44	Carcinoid	IA	64	M	67	Carcinoid	IB	70	M
22	Adenocarcinoma	IB	72	F	45	Epidermoid	IIIA	48	M	68	LCC	IA	73	M
23	Adenosquamous	IA	73	M	46	LCC	IIB	60	M	69	Sarcomatoid	IB	68	M

Table A5 (continued)

Patient no.	Diagnosis	Stage	Age	Gender	Patient no.	Diagnosis	Stage	Age	Gender	Patient no.	Diagnosis	Stage	Age	Gender
70	LCC	IIB	78	M	87	Epidermoid	IB	60	M	104	Adenocarcinoma	IB	74	M
71	Carcinoid	inconclusive	71	M	88	SCC	n.a.	70	M	105	Adenocarcinoma	IB	72	M
72	Adenocarcinoma	IIIB	60	F	89	Epidermoid	IIB	58	M	106	Epidermoid	IIIA	63	M
73	Carcinoid	n.a.	50	F	90	SCC	IIIA	52	M	107	Adenocarcinoma	n.a.	41	F
74	Epidermoid	IIB	64	M	91	Adenocarcinoma	IA	64	M	108	Epidermoid	IA	58	M
75	Adenocarcinoma	IA	71	F	92	LCC	IB	68	M	109	Adenocarcinoma	IIB	59	M
76	Carcinoid	IA	30	F	93	Adenocarcinoma	IB	75	F	110	Adenosquamous	IIB	36	F
77	Sarcomatoid	IA	57	M	94	Epidermoid	IB	62	M	111	Carcinoid	IA	72	M
78	Adenocarcinoma	IB	59	F	95	Adenocarcinoma	n.a.	61	M	112	Adenosquamous	IB	68	M
79	Adenocarcinoma	n.a.	52	F	96	Adenocarcinoma	IIA	56	F	113	Adenocarcinoma	IA	78	F
80	Adenocarcinoma	IA	54	F	97	Epidermoid	IB	63	M	114	Carcinoid	IA	31	F
81	Epidermoid	IB	64	M	98	Carcinoid	IA	55	F	115	Adenosquamous	IB	56	M
82	LCC	IB	54	M	99	Adenocarcinoma	IIIA	66	M	116	Carcinoid	IA	32	M
83	Epidermoid	IA	59	M	100	Epidermoid	IB	61	F	117	Epidermoid	IIIA	46	M
84	Adenocarcinoma	IA	77	M	101	Adenosquamous	IB	70	M	118	Epidermoid	IA	68	M
85	SCC	n.a.	77	M	102	Adenocarcinoma	IIIA	60	M	119	Epidermoid	IA	72	M
86	Adenocarcinoma	n.a.	75	M	103	Adenocarcinoma	IIIA	55	M	120	Epidermoid	IB	48	M

n.a. not available

Table A6 Histological characteristics of lung tumours from which tissue samples were available for HRMAS NMR analysis.

Histology group /subject no.[a]	Tumour size[b] (cm)	TNM	Stage[c]	Tumor cells[d] (%)	Necrosis[d] (%)
Adenocarcinoma (AdC)					
5	3.5	T2N0Mx	IB	80	20
7	3	T2N0Mx	IB	n.a.	n.a.
8	4	T2N1Mx	IIB	40	0
12	3.4	T2N0Mx	IB	80	0
20	3	T2N1Mx	IIB	30	0
25	2.5	T1N0Mx	IA	40	0
28	2.8	T2N0Mx	IB	40	0
33	7	T2N0Mx	IB	90	5
38	3.5	T2N2Mx	IIIA	5	90
48	9	T2N1Mx	IIB	70	5
50	5	T2N0Mx	IB	50	0
52	10	T2N0Mx	IB	40	0
57	4	T2N1Mx	IIB	40	0
64	2.3	T1N0Mx	IA	90	0
65	8	T3N0Mx	IIB	50	0
93	1.9	T2N0Mx	IB	30	0
95	3	n.a.	n.a.	90	0
102	2.2	T1N2Mx	IIIA	30	0
109	5.5	T2N1Mx	IIB	50	0
Squamous cell carcinoma (SqCC)					
19	3	T1N1Mx	IIA	10	20
29	4	T2N1Mx	IIB	30	0
34	5.5	T2N0Mx	IB	5	0
35	9	T2N0Mx	IB	80	0
43	5	T2N1Mx	IIB	20	0
47	n.a.	T2N1Mx	IIB	65	10
54	3	T1N0Mx	IA	40	5
56	6.5	T3N0Mx	IIB	30	30
61	3.5	T2N0Mx	IB	50	0
74	4	T2N1Mx	IIB	50	0
81	4.5	T2N0Mx	IB	10	90
87	1.8	T2N0Mx	IB	50	10
97	4	T2N0Mx	IB	100	0
100	5	T2N0Mx	IB	35	35
106	8	T3N1Mx	IIIA	n.a.	n.a.
117	6	T2N2Mx	IIIA	90	5
118	3.5	T1N0Mx	IA	90	5
119	2.3	T1N0Mx	IA	5	5
120	3.3	T2N0Mx	IB	10	90

Table A6 (continued)

Histology group /subject no.[a]	Tumour size[b] (cm)	TNM	Stage[c]	Tumor cells[d] (%)	Necrosis[d] (%)
Sarcomatoid					
14	3	T1N0Mx	IA	20	0
30	3	T1NxMx	n.a.	50	0
39	4	T2N1Mx	IIB	80	0
42	2.5	T1N0Mx	IA	10	0
66	2	T1N0Mx	IA	60	0
68	4.5	T2N0Mx	IB	80	0
Large cell					
15	5	T2N2Mx	IIIA	10	40
46	5.2	T2N1Mx	IIB	80	0
70	7	T3N0Mx	II	50	50
82	2.7	T2N0Mx	IB	70	0
Adenosquamous					
110	4.5	T2N1Mx	IIA	n.a.	n.a.
115	5	T2N0Mx	IB	55	20
Carcinoid					
4	3.7	T1N0Mx	IA	5	0
44	3	T1N0Mx	IA	90	0
49	2	T1N1Mx	IIA	99	0
111	2.3	T1N0Mx	IA	90	0
Small cell (SCLC)					
55	6	T2N1Mx	IIB	80	0
90	6	T2N2Mx	IIIA	90	1

n.a. not available; [a] subject number according to Table A5; [b]tumour's greatest dimension; [c] stage according to the TNM classification system; [d] obtained by microscopic observation of mirror sections of the tissues analysed by NMR.

ANNEX V: Demographic information on healthy volunteers

Table A7 Demographic and smoking status information on healthy volunteers enrolled in this work.

Control no.	Gender	Age	Smoking status	Control no.	Gender	Age	Smoking status	Control no.	Gender	Age	Smoking status
1	M	52	former smoker	24	M	27	former smoker	47	F	29	non-smoker
2	F	26	non-smoker	25	M	48	former smoker	48	F	49	non-smoker
3	F	44	smoker	26	M	30	smoker	49	F	37	non-smoker
4	F	45	non-smoker	27	M	44	smoker	50	M	56	smoker
5	F	42	former smoker	28	F	44	former smoker	51	F	30	non-smoker
6	F	29	non-smoker	29	F	38	non-smoker	52	F	38	smoker
7	F	51	smoker	30	M	48	non-smoker	53	M	52	non-smoker
8	F	40	smoker	31	M	30	smoker	54	F	48	smoker
9	F	42	non-smoker	32	F	38	smoker	55	M	59	non-smoker
10	F	47	non-smoker	33	F	49	non-smoker	56	F	35	non-smoker
11	M	50	former smoker	34	M	45	non-smoker	57	M	35	smoker
12	M	51	non-smoker	35	F	28	N	58	M	30	former smoker
13	F	52	former smoker	36	F	44	smoker	59	F	47	non-smoker
14	M	33	non-smoker	37	F	37	non-smoker	60	F	22	smoker
15	F	26	non-smoker	38	F	48	non-smoker	61	M	35	smoker
16	F	56	former smoker	39	F	33	smoker	62	F	54	non-smoker
17	M	56	former smoker	40	F	34	smoker	63	M	48	former smoker
18	M	36	smoker	41	F	45	non-smoker	64	M	35	smoker
19	F	49	non-smoker	42	M	25	smoker	65	M	40	non-smoker
20	M	50	non-smoker	43	M	57	smoker	66	M	32	smoker
21	M	46	non-smoker	44	M	38	non-smoker	67	M	23	non-smoker
22	F		non-smoker	45	M	28	smoker	68	M	30	non-smoker
23	M	28	non-smoker	46	M	51	smoker	69	M	43	smoker

Table A7 (continued)

Control no.	Gender	Age	Smoking status	Control no.	Gender	Age	Smoking status	Control no.	Gender	Age	Smoking status
70	M	60	former smoker	79	M	34	non-smoker	88	M	49	non-smoker
71	F	47	non-smoker	80	M	36	non-smoker	89	F	51	former smoker
72	M	45	former smoker	81	F	30	former smoker	90	M	36	non-smoker
73	F	53	former smoker	82	F	54	non-smoker	91	M	48	non-smoker
74	F	48	non-smoker	83	F	54	non-smoker	92	M	43	smoker
75	M	52	former smoker	84	M	53	smoker	93	F	50	non-smoker
76	M	46	former smoker	85	M	58	former smoker	94	F	50	smoker
77	M	45	former smoker	86	F	53	smoker	95	M	40	non-smoker
78	M	26	smoker	87	F	50	former smoker				

ANNEX VI: T_1 and T_2 measurements

Table A8 ^{1}H T_1 and T_2 relaxation time constants measured for selected metabolites of human tumour tissue.

δ (ppm)	Assignment	T_1±SD (ms)	T_2±SD (ms)
0.88	Lipids, -CH_3	208.8±10.8	101.4±5.6
1.00	Isoleucine	439.3±14.6[a]	440.9±77.3[a]
	Broad peak underneath	92.1±3.7	26.9±4.6
1.04	Valine	621.1±19.4[a]	505.2±16.7[a]
	Broad peak underneath	101.0±4.6	52.9±6.5
1.28	Lipids, -CH_2	300.4±10.5	88.6±3.3
1.33	Lactate	934.1±10.4[a]	368.7±10.2[a]
	Broad peak underneath	202.2±10.0	27.6±3.0
1.48	Alanine	888.7±19.5[a]	632.0±15.4 [a]
	Broad peak underneath	145±8.3	19.0±2.3
1.72	Lysine	504.3±20.2[a]	311.8±8.6[a]
	Broad peak underneath	120.3±10.6	33.2±1.2
2.35	Glutamate	828.2±29.2[a]	483.2±11.0[a]
	Broad peak underneath	186.9±16.9	40.3±2.7
2.45	Glutamine	637.5±10.6[a]	463.4±15.2[a]
	Broad peak underneath	99.0±3.5	5.3±0.8
2.56	Glutathione reduced	455.2±17.6[a]	231.8±22.0[a]
	Broad peak underneath	90.1±5.1	6.4±1.2
2.81	Aspartate	556.5±13.3[a]	430.0±43.2[a]
	Broad peak underneath	91.4±4.8	21.1±2.7
2.89	Asparagine	600.2±3.2[a]	[b]
	Broad peak underneath	96.4±4.9	
3.21	Choline	1407.0±47.5[a]	657.4±38.5
	Broad peak underneath	130.5±9.1	
3.22	Phosphocholine	[b]	266.1±12.5
3.23	Glycerophosphocholine	533.8±9.8	436.9±15.3
3.35	*scyllo*-Inositol	1038.5±51.3	417.4±25.0
3.43	Taurine	1563.2±40.3	553.4±13.9
3.56	Glycine	2005.1±22.6	580.8±12.3
3.85	Serine	1568.0±112.8[a]	579.5±13.9[a]
	Broad peak underneath	280.2±18.6	69.6±3.8

Table A8 (continued)

δ (ppm)	Assignment	T_1±SD (ms)	T_2±SD (ms)
3.93	Creatine	755.3±9.8	491.1±4.6[a]
	Broad peak underneath		48.3±2.9
3.97	Phosphoethanolamine	861.7±39.6[a]	547.8±10.2[a]
	Broad peak underneath	164.8±19.4	58.7±3.4
4.06	*myo*-Inositol	781.2±24.7[a]	523.8±23.3[a]
	Broad peak underneath	154.9±9.8	48.3±4.5
4.26	Threonine	1098.0±60.2[a]	352.7±17.0[a]
	Broad peak underneath	231.9±20.7	19.8±1.4
4.65	β-Glucose	689.1±29.3[a]	511.4±22.2[a]
	Broad peak underneath	54.5±6.6	35.7±7.1
5.24	α-Glucose	1342.4±69.8[a]	572.2±58.9[a]
	Broad peak underneath	92.8±7.7	69.5±7.1
5.43	Glycogen	997.5±86.0[a]	236.5±37.5[a]
	Broad peak underneath	145.9±11.7	70.1±20.9
6.88	Tyrosine	1387.1±43.6[a]	462.8±15.8
	Broad peak underneath	123.9±4.4	
7.32	Phenylalanine	1775.2±48.9[a]	688.5±28.8
	Broad peak underneath	97.3±5.1	

[a] Bi-exponential decay was observed and evaluated. [b] Peaks not detected or with low intensity originating poor data were not fitted. SD: fitting standard deviation.

Table A9 ^{1}H T_1 and T_2 relaxation time constants measured for selected metabolites of human blood plasma.

δ (ppm)	Assignment	T_1±SD (ms)	T_2±SD (ms)
0.69	Cholesterol	219.9±3.9	6.0±0.4
0.84	Fatty acyl chains of phospholipids (CH_3, mainly in HDL)	289.0±4.6[a]	4.3±0.4[a]
	Mobile lipids		61.7±3.7
0.87	Fatty acyl chains of triglycerides (CH_3, mainly in LDL+VLDL)	357.6±7.3 [a]	11.9±0.8[a]
	Mobile lipids		180.4±10.0
1.03	Valine	528.3±25.5 [a]	399.8±96.2[a]
	Broad peak underneath	164.0±6.5	6.3±0.5
1.46	Alanine	1205.4±72.5 [a]	1014.2±144.0[a]
	Broad peak underneath	221.1±5.2	12.0±0.7
1.56	Fatty acyl chains in lipoproteins (CH_2-CH_2-CO)	269.0±3.4	16.5±0.9

Table A9 (continued)

δ (ppm)	Assignment	T_1±SD (ms)	T_2±SD (ms)
2.02	N-acetyl groups in glycoproteins	850.8±22.3 [a]	221.9±16.3[a]
	Broad peak underneath	259.0±6.5	18.9±1.4
2.36	Pyruvate	1971.6±94.8 [a]	718.1±138.4[a]
	Broad peak underneath	226.8±3.2	10.8±1.0
2.45	Glutamine	1013.8±53.7 [a]	713.15±249.2[a]
	Broad peak underneath	211.8±3.7	10.0±0.9
3.00	Albumin	241.5±2.6	13.6±0.7
3.03	Creatine	1867.6±80.2 [a]	636.9±82.7[a]
	Broad peak underneath	227.0±3.6	12.7±0.7
3.04	Creatinine	2070.8±89.1 [a]	1377.1±203.8[a]
	Broad peak underneath	233.2±3.7	12.8±0.7
3.21	Choline	435.1±19.2 [a]	149.6±4.9[a]
	Choline of phospholipids in lipoproteins	200.5±12.2	7.5±0.6
3.35	Methanol	2436.5±127.5	713.8±73.7
3.49	β-Glucose	1423.8±20.9 [a]	623.3±19.1[a]
	Broad peak underneath	249.1±4.4	14.0±0.5
3.56	Glycine	2252.2±89.2	732.2±69.9
4.11	Lactate	2038.7±57.9 [a]	749.9±20.6[a]
	Broad peak underneath	264.6±3.8	9.0±0.1
5.23	α-Glucose	2136.7±298.6 [a]	397.5±37.5[a]
	Broad peak underneath	641.1±132.6	14.6±2.3
5.26	Fatty acyl chains of phospholipids (CH=CH, mainly in HDL)	507.0±10.0 [a]	27.9±2.3
5.30	Fatty acyl chains of triglycerides (CH=CH, mainly in LDL+VLDL)	592.6±9.1	29.7±4.5
7.72	Protein	269.6±6.6	3.2±0.6

[a] Bi-exponential decay was observed and evaluated. SD: fitting standard deviation.

Table A10 ^{1}H T_1 and T_2 relaxation time constants measured for selected metabolites of human urine.

δ (ppm)	Assignment	T1±SD (ms)	δ (ppm)	Assignment	T1±SD (ms)
2.24[a]	Acetone	3.378±0.327	1.36	α-Hydroxyisobutyrate	1230.1±20.3
1.99	N-acetylated metabolite	1568.1±44.5	1.27	β-Hydroxyisovalerate	967.9±12.2
2.01	N-acetylated metabolite	1501.0±68.0	6.98	p-Hydroxyhippurate	2.738±0.083
2.03	N-acetylated metabolite	1165.2±26.0	6.87	*p*-Hydroxyphenylacetate	3.406±0.104

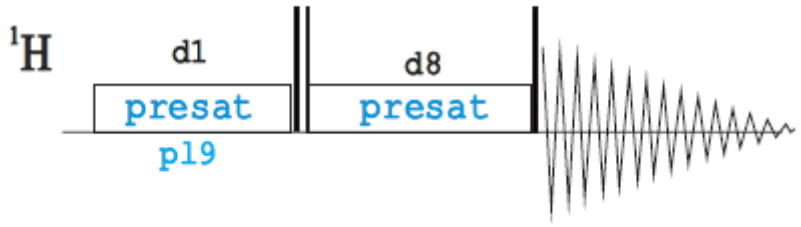
Table A10 (continued)

δ (ppm)	Assignment	T1±SD (ms)	δ (ppm)	Assignment	T1±SD (ms)
2.05	N-acetylated metabolite	1067.8±20.9	7.71	Indoxyl sulfate	6.110±0.202
2.07	N-acetylated metabolite	1062.5±21.4	3.36	*scyllo*-Inositol	1.373±0.009
2.08	N-acetylated metabolite	0.886±0.012	1.33	Lactate	951.8±15.8
1.48	Alanine	973.6±19.5	7.06	1-Methylhistidine	2.812±0.050
2.56	Citrate	0.764±0.003	7.15	3- Methylhistidine	3.180±0.048
2.68	Citrate	0.772±0.003	7.43	Phenylacetylglutamine	3.831±0.047
3.05	Creatinine	2.824±0.023	4.35	Tartrate	2.862±0.064
4.06	Creatinine	2.274±0.011	3.43	Taurine	1.988±0.018
2.36	*p*-Cresol	2.092±0.030	1.33	Threonine	748.8±7..5
1.23	4-DEA	0.841±0.013	3.27	TMAO	2.318±0.006
2.73	Dimethylamine	3.694±0.017	4.44	Trigonelline	1.861±0.068
1.11	4-DTA	777.8±16.6	1.07	Unknown	940.6±25.0
1.55	Fatty acids	0.531±0.010	1.15	Unknown	0.643±0.017
8.46	Formate	10.195±0.166	2.46	Unknown	1.177±0.014
3.57	Glycine	2.218±0.035	4.10	Unknown	2.101±0.039
7.56	Hippurate	3.196±0.014	4.30	Unknown	1.433±0.034
7.64	Hippurate	4.561±0.043	5.35	Unknown	1.238±0.294
7.84	Hippurate	3.220±0.010	6.67	Unknown	7.411±0.775
2.42	β-Hydroxybutyrate	4.468±0.228 [a]	7.68	Unknown	1.902±0.014
	Broad peak underneath	0.718±0.038	2.78	Unknown	1.324±0.016

[a] Bi-exponential decay was observed and evaluated.

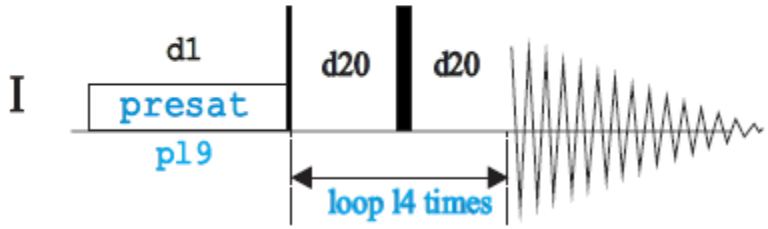
ANNEX VII: Schematic representation of the 1D ^{1}H NMR pulse programmes used

Pulse programme *noesypr1d* (standard 1D) from Bruker library[1]:



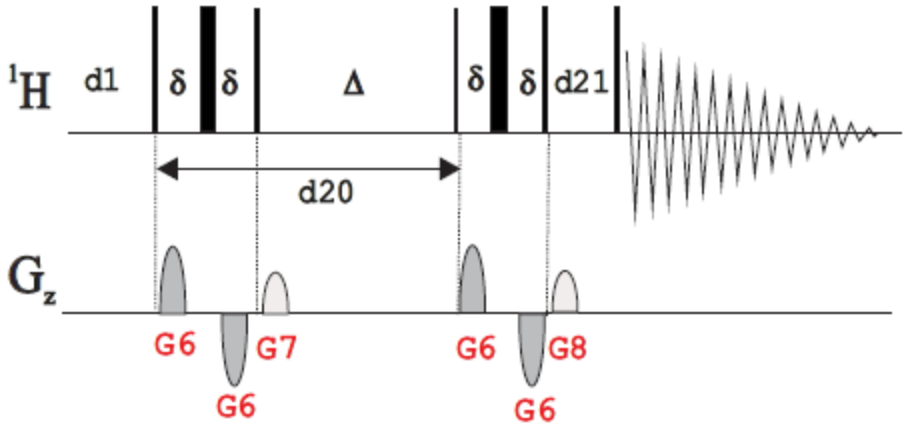
pl9: power level for presaturation; d1: relaxation delay; d8: mixing time.

Pulse programme *cpmgpr* (CPMG) from Bruker library[1]:



pl9: power level for presaturation; d1: relaxation delay; d20: fixed echo time; l4: loop for T_2 filter.

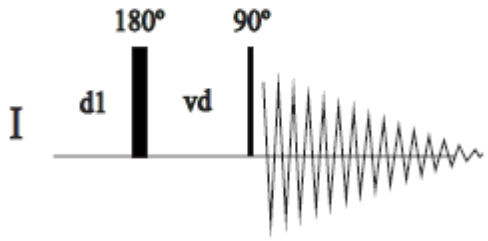
Pulse programme *ledbpgp2s1d* (diffusion-edited) from Bruker library[1]:



d1: relaxation delay; d20: diffusion time; d21 eddy current delay; G6, G7, G8: *z*-gradients. A pulse for presaturation (with pl9 power) is also applied during d1 (not shown in the scheme)

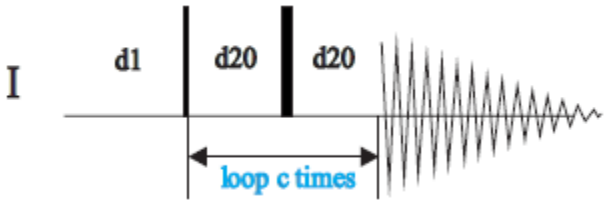
[1] Parella , T., 2006. Pulse programme catalogue: volume I: 1D & 2D NMR experiments. *NMR guide 4.1 – Topspin 2.0*. Germany, Bruker Biospin GmbH.

Pulse programme *t1irpr* (T_1 measurement) from Bruker library[1]:



d1: relaxation delay; vd: variable delay obtained from the vd list. A pulse for presaturation (with pl9 power) is also applied during d1 (not shown in the scheme).

Pulse programme *t2zgpr* (T_2 measurement) from Bruker library[1]:



d1: relaxation delay; d20: fixed echo time; c: counter the number of loops obtained from the vc list. A pulse for presaturation (with pl9 power) is also applied during d1 (not shown in the scheme).

[1] Parella , T., 2006. Pulse programme catalogue: volume I: 1D & 2D NMR experiments. *NMR guide 4.1 – Topspin 2.0*. Germany, Bruker Biospin GmbH.

ANNEX VIII: List of UPLC-MS features relevant to cancer vs. control discrimination

Table A11 List of significant UPLC-MS HSS ESI+ features (Wilcoxon rank sum test, $p<0.01$) obtained for the comparison of the cancer group compared to controls. Features intensities are indicated as: very high ($>10^5$), high ($>10^4$, $<10^5$), low ($>10^3$, $<10^4$) and very low ($<10^3$).

Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI	Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI
1	0.43	520.77	very low	82.3±14.6	1.1×10^{-4}	0.82±0.41	22	1.00	474.04	very low	51.6±5.2	3.9×10^{-12}	1.61±0.46
2	0.43	316.87	very low	105.3±16.5	1.9×10^{-4}	0.86±0.41	23	1.00	315.05	very low	44.6±8.3	1.0×10^{-4}	0.90±0.42
3	0.44	248.88	very low	55.0±11.4	8.1×10^{-4}	0.77±0.41	24	1.00	576.09	very low	54.7±10.3	1.3×10^{-4}	0.86±0.41
4	0.46	456.71	very low	55.7±10.7	2.0×10^{-4}	0.83±0.41	25	1.00	458.07	low	43.5±9.0	3.3×10^{-4}	0.82±0.41
5	0.47	256.82	very low	61.6±12.1	1.0×10^{-4}	0.80±0.41	26	1.00	442.09	very low	44.5±6.8	6.4×10^{-7}	1.10±0.43
6	0.50	104.11	very low	26.5±6.0	5.6×10^{-4}	0.79±0.41	27	1.00	155.04	very low	17.8±3.9	2.1×10^{-4}	0.85±0.41
7	0.50	257.00	very low	43.7±6.4	2.6×10^{-7}	1.15±0.43	28	1.00	203.55	very low	35.7±3.7	7.5×10^{-12}	1.69±0.46
8	0.55	114.07	high	22.1±3.4	8.2×10^{-8}	1.18±0.43	29	1.00	245.08	low	89.8±15.0	8.6×10^{-7}	0.85±0.41
9	0.57	139.06	low	-59.6±14.9	7.6×10^{-8}	-1.13±0.43	30	1.00	513.11	low	68.3±7.5	3.4×10^{-9}	1.40±0.44
10	0.65	335.10	low	86.9±10.4	1.0×10^{-7}	1.20±0.43	31	1.00	386.58	low	53.6±5.0	8.8×10^{-14}	1.72±0.46
11	0.85	285.23	low	66.4±8.3	2.6×10^{-7}	1.22±0.43	32	1.00	573.11	low	57.2±5.3	1.3×10^{-13}	1.72±0.46
12	0.86	203.03	low	39.2±4.2	2.4×10^{-10}	1.58±0.45	33	1.00	386.08	low	55.8±6.6	1.3×10^{-8}	1.36±0.44
13	0.99	460.08	very low	57.4±9.9	3.9×10^{-5}	0.93±0.42	34	1.00	345.00	very low	69.0±10.3	5.5×10^{-6}	1.02±0.42
14	0.99	219.14	low	-44.3±10.4	4.9×10^{-8}	-1.09±0.42	35	1.00	264.55	very low	56.2±6.8	2.9×10^{-8}	1.33±0.44
15	0.99	285.57	very low	51.6±5.2	3.9×10^{-12}	1.61±0.46	36	1.00	264.05	very low	56.5±7.0	2.0×10^{-8}	1.30±0.44
16	1.00	444.09	very low	33.7±4.2	7.5×10^{-12}	1.41±0.44	37	1.00	557.12	very low	74.3±7.5	4.3×10^{-10}	1.50±0.45
17	1.00	575.12	very low	57.0±9.4	3.5×10^{-5}	0.97±0.42	38	1.01	543.06	very low	59.6±6.2	1.3×10^{-10}	1.53±0.45
18	1.00	515.12	very low	49.0±8.5	3.4×10^{-5}	0.94±0.42	39	1.01	373.08	very low	28.4±7.9	1.3×10^{-3}	0.65±0.41
19	1.00	577.09	very low	51.6±7.2	5.5×10^{-7}	1.17±0.43	40	1.01	284.56	low	58.2±8.8	5.0×10^{-6}	1.06±0.42
20	1.00	387.09	low	52.2±6.2	8.1×10^{-9}	1.36±0.44	41	1.01	313.06	low	31.0±6.3	5.3×10^{-5}	0.88±0.41
21	1.00	387.09	low	35.3±6.8	1.9×10^{-5}	0.91±0.42	42	1.01	289.04	very low	66.6±12.7	5.2×10^{-4}	0.81±0.41

Table A11 (continued)

Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI	Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI
43	1.01	405.05	very low	76.1±10.1	4.6×10^{-6}	1.13±0.43	69	1.98	448.06	low	-60.3±17.3	5.7×10^{-9}	-0.98±0.42
44	1.01	308.09	very low	33.9±6.0	2.0×10^{-5}	0.99±0.42	70	1.99	171.06	very low	-64.8±23.6	1.1×10^{-6}	-0.80±0.41
45	1.01	347.05	very low	51.5±12.6	1.1×10^{-3}	0.66±0.41	71	1.99	341.01	low	46.5±7.2	2.7×10^{-6}	1.07±0.42
46	1.01	329.03	low	70.3±10.5	5.5×10^{-6}	1.02±0.42	72	2.18	361.21	very low	-64.1±17.9	2.5×10^{-10}	-1.03±0.42
47	1.01	370.05	very low	60.3±14.8	5.2×10^{-3}	0.64±0.41	73	2.40	185.09	very low	-46.1±14.4	2.0×10^{-6}	-0.83±0.41
48	1.09	241.12	very low	-56.5±15.1	6.8×10^{-8}	-1.04±0.42	74	2.42	424.25	very low	121.8±16.9	1.2×10^{-4}	0.92±0.42
49	1.15	438.98	low	59.6±11.5	6.4×10^{-5}	0.81±0.41	75	2.44	347.16	very low	-75.3±27.7	3.4×10^{-9}	-0.86±0.41
50	1.15	308.01	low	85.4±17.1	5.0×10^{-5}	0.72±0.41	76	2.47	343.06	low	-81.5±24.8	2.6×10^{-10}	-1.09±0.42
51	1.15	232.52	very low	-70.8±37.6	1.3×10^{-5}	-0.57±0.40	77	2.49	410.04	very low	99.6±11.7	2.0×10^{-7}	1.17±0.43
52	1.23	169.04	high	48.4±8.8	5.8×10^{-5}	0.90±0.42	78	2.49	470.07	very low	90.7±13.7	1.7×10^{-5}	0.94±0.42
53	1.24	393.02	very low	71.5±12.8	3.8×10^{-5}	0.84±0.41	79	2.49	451.06	very low	139.1±17.3	9.9×10^{-7}	0.98±0.42
54	1.25	417.03	very low	86.4±15.6	4.7×10^{-5}	0.79±0.41	80	2.52	284.15	low	143.9±15.9	2.0×10^{-7}	1.08±0.42
55	1.25	604.98	very low	71.5±16.9	9.2×10^{-4}	0.64±0.41	81	2.52	279.10	low	93.2±8.2	2.4×10^{-10}	1.59±0.45
56	1.26	368.16	very low	-64.0±16.6	7.2×10^{-12}	-1.12±0.43	82	2.52	182.08	low	78.7±8.5	6.3×10^{-9}	1.36±0.44
57	1.27	261.09	low	846.3±22.0	9.4×10^{-13}	1.54±0.45	83	2.53	425.20	very low	-69.1±24.4	2.3×10^{-11}	-0.85±0.41
58	1.27	323.06	very low	1823.8±26.0	6.4×10^{-14}	1.45±0.45	84	2.56	306.03	very low	72.5±9.9	2.3×10^{-6}	1.10±0.42
59	1.32	296.05	very low	89.7±14.2	6.7×10^{-6}	0.90±0.42	85	2.58	508.14	very low	-59.9±30.1	4.3×10^{-10}	-0.57±0.40
60	1.33	255.03	very low	88.5±13.5	4.7×10^{-5}	0.93±0.42	86	2.59	275.11	low	-70.4±38.0	4.2×10^{-8}	-0.56±0.40
61	1.34	302.09	low	-71.2±31.1	3.2×10^{-7}	-0.70±0.41	87	2.60	441.15	very low	126.9±14.2	5.3×10^{-7}	1.13±0.43
62	1.59	317.08	very low	-60.1±20.2	2.5×10^{-6}	-0.84±0.41	88	2.65	302.16	very low	-53.3±14.1	8.9×10^{-7}	-1.01±0.42
63	1.63	293.05	low	74.9±11.6	1.8×10^{-5}	0.96±0.42	89	2.70	256.12	low	-73.9±33.6	9.2×10^{-7}	-0.68±0.41
64	1.82	272.13	low	148.6±21.2	5.7×10^{-9}	0.84±0.41	90	2.70	321.06	low	36.0±6.8	6.0×10^{-5}	0.92±0.42
65	1.90	221.54	very low	91.1±10.9	9.6×10^{-8}	1.18±0.43	91	2.70	346.11	low	34.1±6.7	3.1×10^{-5}	0.89±0.42
66	1.91	401.05	low	89.5±10.8	2.2×10^{-7}	1.17±0.43	92	2.73	384.18	low	-83.3±47.1	8.4×10^{-9}	-0.59±0.40
67	1.96	170.09	low	-33.0±17.6	5.0×10^{-4}	-0.45±0.40	93	2.77	303.07	very low	-86.0±74.0	$2.1\pm\times10^{-9}$	-0.40±0.40
68	1.96	508.08	very low	-63.9±25.6	1.3×10^{-6}	-0.72±0.41	94	2.77	375.22	low	-74.2±23.5	2.4×10^{-12}	-0.98±0.42

Table A11 (continued)

Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI	Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI
95	2.81	285.14	very low	-74.6±28.2	4.3×10^{-9}	-0.83±0.41	107	3.42	244.15	high	-30.6±7.0	1.1×10^{-6}	-1.03±0.42
96	2.85	369.08	very low	-61.6±18.1	1.3×10^{-6}	-0.98±0.42	108	3.87	522.09	very low	136.1±14.7	4.6×10^{-9}	1.14±0.43
97	2.95	334.11	high	27.6±4.6	5.0×10^{-7}	1.08±0.42	109	3.88	341.18	low	-68.4±20.1	5.8×10^{-11}	-1.02±0.42
98	2.96	375.14	low	34.7±6.7	2.1×10^{-5}	0.90±0.42	110	3.88	764.21	very low	42.2±9.9	1.2×10^{-7}	0.71±0.41
99	2.99	207.02	very low	-89.6±40.2	1.1×10^{-7}	-0.79±0.41	111	3.88	171.08	low	46.9±6.0	4.5×10^{-8}	1.30±0.44
100	3.01	409.21	very low	-54.7±22.0	9.3×10^{-8}	-0.67±0.41	112	3.88	780.19	low	122.5±12.6	5.1×10^{-9}	1.24±0.43
101	3.05	277.10	low	-56.0±12.0	8.1×10^{-10}	-1.28±0.43	113	4.35	319.13	very low	-73.0±16.9	1.7×10^{-12}	-1.33±0.44
102	3.05	237.12	low	-72.5±17.1	6.4×10^{-10}	-1.31±0.44	114	4.67	289.11	low	-85.8±26.5	5.3×10^{-12}	-1.11±0.43
103	3.05	237.12	low	-69.1±17.2	2.7×10^{-9}	-1.20±0.43	115	4.68	267.12	low	-72.5±20.8	8.8×10^{-9}	-1.08±0.42
104	3.05	175.12	low	-67.6±16.1	3.3×10^{-9}	-1.25±0.43	116	5.32	244.19	low	-63.2±24.0	7.3×10^{-8}	-0.77±0.41
105	3.05	255.13	low	-72.5±17.1	6.4×10^{-10}	-1.31±0.44	117	6.95	409.16	low	572.5±13.6	2.1×10^{-15}	2.27±0.51
106	3.40	343.10	low	-68.4±14.3	6.2×10^{-12}	-1.43±0.44	118	6.96	425.14	low	641.9±14.9	5.2×10^{-15}	2.14±0.50

Table A12 List of assigned significant UPLC-MS HSS ESI- features (Wilcoxon rank sum test, p<0.01) obtained for the comparison of the cancer group compared to controls. Features intensities are indicated as: very high (>10^5), high (>10^4, <10^5), low (>10^3, <10^4) and very low (<10^3).

Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI	Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI
1	1.68	513.99	very low	68.5±14.5	1.9×10^{-4}	0.72±0.41	9	1.87	561.96	very low	54.9±8.0	4.0×10^{-7}	1.10±0.42
2	1.68	452.02	very low	78.2±19.3	4.8×10-^{4}	0.60±0.40	10	1.88	515.96	very low	78.3±8.6	1.0×10^{-8}	1.34±0.44
3	1.76	401.06	low	60.6±9.3	3.2×10-^{6}	1.02±0.42	11	1.88	462.94	very low	60.8±7.8	3.6×10^{-8}	1.24±0.43
4	1.76	337.93	very low	56.8±9.3	1.2×10^{-5}	0.97±0.42	12	1.88	499.99	low	47.2±7.7	1.9×10^{-5}	1.02±0.42
5	1.81	408.95	low	-82.1±36.5	7.9×10^{-8}	-0.75±0.41	13	1.88	583.94	very low	85.6±11.1	1.1×10^{-6}	1.11±0.43
6	1.82	192.98	high	-54.9±17.6	6.2×10^{-7}	-0.85±0.41	14	1.94	417.98	low	71.7±14.1	5.6×10^{-4}	0.77±0.41
7	1.87	410.03	high	35.4±13.3	1.1×10^{-7}	0.47±0.40	15	1.95	547.93	very low	108.0±21.0	2.4×10^{-3}	0.69±0.41
8	1.87	493.98	low	49.4±7.9	4.2×10^{-6}	1.02±0.42	16	1.95	196.05	low	85.0±8.4	3.3×10^{-9}	1.45±0.45

Table A12 (continued)

Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI	Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI
17	1.96	204.98	low	-73.7±23.2	1.1×10^{-8}	-0.99±0.42	43	2.5	460.99	low	-72.3±23.5	2.9×10^{-6}	-0.95±0.42
18	1.96	171.03	low	52.2±11.9	1.1×10^{-3}	0.71±0.41	44	2.5	219	high	-63.4±13.6	6.0×10^{-9}	-1.35±0.44
19	1.97	537.98	very low	63.7±8.2	6.0×10^{-7}	1.20±0.43	45	2.52	204	very low	-60.2±16.3	2.8×10^{-7}	-1.05±0.42
20	1.97	185.96	very low	55.7±8.9	1.7×10^{-5}	1.00±0.42	46	2.52	310.01	very low	331.2±17.2	2.9×10^{-10}	1.50±0.45
21	1.97	233	low	53.3±10.6	7.4×10^{-5}	0.81±0.41	47	2.52	361.11	low	177.1±10.9	9.8×10^{-11}	1.78±0.47
22	1.97	229.07	low	95.4±13.8	8.8×10^{-6}	0.96±0.42	48	2.52	267.13	low	166.5±10.1	1.1×10^{-11}	1.85±0.47
23	1.99	479.95	low	81.9±14.5	8.2×10^{-6}	0.82±0.41	49	2.52	196.05	low	177.2±11.8	7.7×10^{-11}	1.65±0.46
24	2.39	282.08	very low	67.3±7.6	1.1×10^{-8}	1.35±0.44	50	2.52	153.04	very low	142.8±21.9	3.0×10^{-8}	0.79±0.41
25	2.4	287.07	low	63.2±11.1	3.0×10^{-5}	0.89±0.42	51	2.52	180.05	low	121.8±8.0	1.4×10^{-12}	1.93±0.48
26	2.4	296.14	very low	197.1±13.2	1.6×10^{-11}	1.55±0.45	52	2.52	248.04	low	162.9±13.8	3.9×10^{-9}	1.34±0.44
27	2.42	312.09	low	73.9±10.8	3.9×10^{-6}	1.03±0.42	53	2.53	302.1	very low	-73.0±26.2	1.4×10^{-8}	-0.86±0.41
28	2.44	212.05	low	73.4±15.2	6.4×10^{-4}	0.73±0.41	54	2.55	305.06	low	45.7±6.4	2.5×10^{-6}	1.19±0.43
29	2.44	406.01	very low	43.3±7.7	1.6×10^{-5}	0.94±0.42	55	2.55	414	very low	90.2±9.8	7.9×10^{-8}	1.30±0.44
30	2.44	342.02	very low	106.6±17.5	1.1×10^{-4}	0.82±0.41	56	2.55	269.08	low	53.3±7.5	1.4×10^{-7}	1.14±0.43
31	2.45	341.07	low	-57.4±16.2	1.3×10^{-6}	-0.98±0.42	57	2.55	447.1	low	35.3±5.9	2.6×10^{-6}	1.04±0.42
32	2.48	523.09	very low	-78.8±24.1	6.2×10^{-7}	-1.06±0.42	58	2.55	377.03	low	52.4±9.2	7.4×10^{-6}	0.92±0.42
33	2.48	587.15	very low	-62.3±34.2	8.4×10^{-5}	-0.52±0.40	59	2.6	351.06	low	75.4±7.8	3.2×10^{-9}	1.43±0.44
34	2.48	590.05	very low	87.5±17.5	3.6×10^{-4}	0.72±0.41	60	2.6	435.01	very low	123.7±11.5	1.4×10^{-9}	1.37±0.44
35	2.48	529.1	low	81.8±13.1	9.1×10^{-6}	0.91±0.42	61	2.6	331.03	very low	168.0±18.1	1.8×10^{-10}	1.05±0.42
36	2.49	423.03	low	43.5±14.3	5.3×10^{-7}	0.51±0.40	62	2.66	433.12	low	55.8±6.9	2.4×10^{-9}	1.30±0.44
37	2.49	469	very low	127.1±17.0	8.2×10^{-6}	0.95±0.42	63	2.67	355.13	very low	162.4±19.5	7.7×10^{-7}	0.95±0.42
38	2.49	739.12	very low	231.4±25.1	3.2×10^{-6}	0.89±0.42	64	2.67	239.08	low	64.6±9.1	1.1×10^{-6}	1.09±0.42
39	2.49	454.03	very low	179.4±20.6	5.1×10^{-7}	0.95±0.42	65	2.67	314.06	very low	-55.3±24.7	3.8×10^{-6}	-0.62±0.41
40	2.49	392.06	low	131.1±15.5	2.5×10^{-7}	1.05±0.42	66	2.67	402.99	low	-83.7±28.4	3.1×10^{-11}	-1.00±0.42
41	2.49	490.04	very low	104.6±13.6	3.2×10^{-6}	1.04±0.42	67	2.72	204.98	high	-79.1±21.6	1.3×10^{-9}	-1.19±0.43
42	2.49	518.98	very low	135.8±19.3	3.3×10^{-5}	0.87±0.41	68	2.73	455.97	very low	-82.7±27.8	1.6×10^{-8}	-0.99±0.42

Table A12 (continued)

Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI	Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI
69	2.74	491.15	very low	-88.6±49.3	4.1×10^{-9}	-0.63±0.41	92	3.28	331.06	low	55.6±6.0	8.8×10^{-10}	1.48±0.45
70	2.78	301.05	low	-74.5±42.6	5.3×10^{-6}	-0.55±0.40	93	3.4	319.1	low	-55.8±10.8	1.1×10^{-10}	-1.42±0.44
71	2.85	382.02	very low	-60.0±22.4	1.6×10^{-7}	-0.76±0.41	94	3.72	303.02	low	-76.9±21.0	8.4×10^{-10}	-1.17±0.43
72	2.86	477.1	very low	-46.0±17.3	1.4×10^{-6}	-0.70±0.41	95	3.82	277.02	low	-67.0±25.8	1.6×10^{-7}	-0.77±0.41
73	2.86	493.01	low	-27.9±18.9	3.0×10^{-8}	-0.34±0.40	96	3.86	178.05	very high	-31.1±6.7	3.8×10^{-7}	-1.11±0.43
74	2.86	573.05	very low	-80.1±42.6	3.8×10^{-7}	-0.61±0.41	97	3.87	443.17	low	-57.5±12.5	7.9×10^{-8}	-1.28±0.43
75	2.91	546.1	very low	-69.9±27.3	9.9×10^{-7}	-0.77±0.41	98	3.87	274.01	low	-90.0±24.5	8.7×10^{-12}	-1.30±0.44
76	2.91	124.04	very low	-76.4±40.3	3.0×10^{-8}	-0.60±0.40	99	3.87	580.14	low	49.1±11.3	6.1×10^{-4}	0.72±0.41
77	2.94	218.1	high	41.8±11.9	2.8×10^{-3}	0.60±0.40	100	3.87	273.01	high	-87.9±23.0	3.7×10^{-12}	-1.33±0.44
78	2.94	354.08	very low	98.9±14.9	5.3×10^{-5}	0.91±0.42	101	3.87	615.21	low	66.1±10.0	5.9×10^{-6}	1.02±0.42
79	2.95	378.1	low	78.7±10.1	5.7×10^{-9}	1.15±0.43	102	3.87	626.14	low	55.0±8.9	4.6×10^{-6}	0.99±0.42
80	2.95	505.07	very low	-86.5±33.6	2.1×10^{-10}	-0.89±0.42	103	3.87	861.3	low	68.1±10.1	5.3×10^{-6}	1.03±0.42
81	2.96	445.1	very low	-79.0±27.4	5.8×10^{-8}	-0.93±0.42	104	3.87	245.1	low	69.7±12.5	1.5×10^{-4}	0.85±0.41
82	2.96	203	low	-85.7±21.2	1.2×10^{-16}	-1.38±0.44	105	3.87	844.25	low	66.9±10.2	2.4×10^{-6}	1.01±0.42
83	2.99	403.03	very low	-75.8±22.3	1.0×10^{-8}	-1.08±0.42	106	3.88	835.3	low	46.3±8.5	1.3×10^{-5}	0.91±0.42
84	2.99	307.05	very low	-81.4±32.0	1.3×10^{-7}	-0.84±0.41	107	3.88	829.27	low	37.9±6.7	3.4×10^{-6}	0.98±0.42
85	2.99	220	low	-59.1±21.7	1.1×10^{-5}	-0.77±0.41	108	3.88	363.03	high	55.7±8.0	2.8×10^{-7}	1.11±0.43
86	2.99	218.99	high	-77.0±24.6	2.0×10^{-9}	-1.00±0.42	109	3.88	813.31	low	-68.9±18.2	3.4×10^{-6}	-1.13±0.43
87	2.99	218.99	high	-75.8±22.3	1.0×10^{-8}	-1.08±0.42	110	3.88	611.18	low	62.0±11.8	4.9×10^{-5}	0.83±0.41
88	3	283.11	low	110.7±14.2	1.5×10^{-6}	1.03±0.42	111	3.89	784.2	low	103.8±15.0	5.0×10^{-7}	0.94±0.42
89	3.05	262.03	low	67.0±9.8	1.6×10^{-8}	1.05±0.42	112	3.89	498.09	high	75.3±10.1	1.1×10^{-7}	1.11±0.43
90	3.08	295.99	low	163.0±15.9	2.7×10^{-8}	1.17±0.43	113	3.91	331.1	low	116.9±11.9	5.7×10^{-9}	1.28±0.43
91	3.1	416.94	low	-79.6±19.0	8.3×10^{-12}	-1.36±0.44							

Table A13 List of assigned significant UPLC-MS HILIC ESI+ features (Wilcoxon rank sum test, $p<0.01$) obtained for the comparison of the cancer group compared to controls. Features intensities are indicated as: very high ($>10^5$), high ($>10^4$, $<10^5$), low ($>10^3$, $<10^4$) and very low ($<10^3$).

Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI	Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI
1	0.44	643.02	very low	-75.9±27.8	2.23×10^{-9}	-0.87±0.41	26	4.16	449.17	very low	-66.4±21.5	5.54×10^{-8}	-0.91±0.42
2	0.46	315.04	very low	-29.0±25.9	8.29×10^{-3}	-0.26±0.40	27	4.21	463.15	low	-50.8±15.8	1.17×10^{-11}	-0.87±0.41
3	0.46	274.14	high	378.7±19.6	8.31×10^{-12}	1.39±0.44	28	4.21	591.08	very low	-48.7±11.0	1.28×10^{-8}	-1.16±0.43
4	0.48	299.01	low	119.0±14.0	1.32×10^{-7}	1.10±0.42	29	4.29	365.02	low	-71.5±19.4	1.50×10^{-9}	-1.13±0.43
5	0.49	369.12	very low	-55.1±20.8	5.34×10^{-7}	-0.72±0.41	30	4.3	327.07	low	-69.5±20.2	1.37×10^{-7}	-1.04±0.42
6	0.59	387.18	low	2055.7±135.0	1.42×10^{-6}	0.28±0.40	31	4.59	406.13	low	63.0±7.6	4.19×10^{-8}	1.29±0.44
7	0.79	392.04	low	-69.4±35.1	4.29×10^{-10}	-0.60±0.40	32	4.59	389.13	low	41.1±5.7	2.62×10^{-7}	1.22±0.43
8	0.79	351.01	very low	-74.1±27.9	8.47×10^{-11}	-0.83±0.41	33	4.59	403.16	very low	-61.4±19.5	1.11×10^{-6}	-0.89±0.42
9	2.47	444.13	very low	-57.3±12.2	4.26×10^{-9}	-1.30±0.44	34	4.77	319.12	low	-65.2±20.2	6.24×10^{-8}	-0.95±0.42
10	2.49	206.05	low	-71.7±22.1	5.88×10^{-10}	-0.99±0.42	35	5	246.01	very low	120.5±16.9	1.28×10^{-6}	0.92±0.42
11	2.5	372.1	low	-57.4±10.5	3.26×10^{-10}	-1.52±0.45	36	5.19	341.17	low	-56.8±11.8	2.13×10^{-9}	-1.33±0.44
12	2.78	457.1	very low	-61.4±12.7	2.48×10^{-10}	-1.38±0.44	37	5.2	177.02	very low	-65.7±18.4	3.43×10^{-11}	-1.04±0.42
13	2.78	532.11	very low	-51.1±10.5	2.23×10^{-9}	-1.29±0.44	38	5.2	194.05	very low	-60.2±16.6	3.91×10^{-9}	-1.02±0.42
14	2.79	180.07	low	-53.4±13.1	4.91×10^{-10}	-1.10±0.42	39	5.29	187.11	low	-41.0±8.2	3.59×10^{-9}	-1.24±0.43
15	3.73	881.28	low	140.6±24.0	5.85×10^{-5}	0.71±0.41	40	5.83	417.2	low	88.9±8.6	3.11×10^{-11}	1.47±0.45
16	3.73	595.18	high	83.1±18.5	2.05×10^{-4}	0.65±0.41	41	5.83	399.19	low	103.6±11.5	3.96×10^{-11}	1.23±0.43
17	3.74	897.25	very low	132.3±22.8	1.04×10^{-5}	0.72±0.41	42	5.94	191.07	high	-41.2±9.6	1.93×10^{-7}	-1.06±0.42
18	3.75	859.29	very low	136.9±19.0	3.66×10^{-6}	0.88±0.41	43	6.12	251.15	low	63.6±15.0	5.18×10^{-5}	0.66±0.41
19	3.86	509.14	very low	-65.4±17.0	1.33×10^{-8}	-1.12±0.43	44	6.3	317.17	very low	-77.7±26.4	3.02×10^{-9}	-0.95±0.42
20	3.86	331.01	very low	-54.5±10.4	9.18×10^{-12}	-1.41±0.44	45	6.34	384.19	low	-76.6±38.4	2.33×10^{-9}	-0.63±0.41
21	3.89	475.04	low	-81.5±37.8	1.02×10^{-10}	-0.71±0.41	46	6.38	133.1	very low	35.0±4.3	3.37×10^{-4}	1.40±0.44
22	3.9	477.09	low	-76.0±23.5	2.59×10^{-10}	-1.02±0.42	47	6.4	203.42	high	36.5±5.4	8.91×10^{-7}	1.16±0.43
23	3.91	437.07	low	-77.8±27.1	1.75×10^{-12}	-0.92±0.42	48	6.41	158.13	low	23.1±3.5	1.15×10^{-6}	1.19±0.43
24	3.92	364.1	low	45.4±6.5	2.52×10^{-7}	1.17±0.43	49	6.57	320.22	very low	120.2±12.5	4.53×10^{-8}	1.22±0.43
25	4.14	272.14	very low	105.2±13.7	3.05×10^{-7}	1.04±0.42	50	6.74	410.17	low	-49.4±18.3	4.80×10^{-6}	-0.71±0.41

Table A13 (continued)

Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI	Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI
51	6.74	171.15	high	38.9±6.3	9.24×10^{-7}	1.07±0.42	59	7.16	421.18	very low	-58.9±15.7	2.37×10^{-8}	-1.06±0.42
52	6.82	419.17	very low	214.1±19.7	1.51×10^{-8}	1.08±0.42	60	7.17	225.12	high	-54.5±14.3	2.62×10^{-7}	-1.03±0.42
53	6.83	271.1	low	-62.6±25.7	2.83×10^{-7}	-0.70±0.41	61	7.27	189.42	low	-33.9±6.3	2.58×10^{-8}	-1.29±0.44
54	6.9	275.21	very low	71.7±9.6	2.01×10^{-7}	1.13±0.43	62	7.42	282.18	low	217.5±14.5	1.74×10^{-11}	1.49±0.45
55	6.91	221.15	very low	72.9±7.3	3.59×10^{-11}	1.51±0.45	63	7.42	246.18	high	56.1±5.8	4.57×10^{-11}	1.54±0.45
56	7.02	328.13	low	81.8±19.0	1.06×10^{-4}	0.63±0.41	64	7.42	268.16	low	378.1±30.3	2.91×10^{-8}	0.90±0.42
57	7.14	589.3	very low	229.3±15.5	2.11×10^{-11}	1.42±0.44	65	7.94	395.21	very low	135.8±16.0	6.42×10^{-7}	1.05±0.42
58	7.16	368.16	low	-68.3±15.6	7.16×10^{-12}	-1.30±0.44	66	7.94	379.23	low	142.9±15.1	4.27×10^{-7}	1.14±0.43

Table A14 List of assigned significant UPLC-MS HILIC ESI- features (Wilcoxon rank sum test, $p<0.01$) obtained for the comparison of the cancer group compared to controls. Features intensities are indicated as: very high ($>10^5$), high ($>10^4$, $<10^5$), low ($>10^3$, $<10^4$) and very low ($<10^3$).

Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI	Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI
1	0.44	597.03	low	-84.1±37.5	9.2×10^{-10}	-0.76±0.41	13	0.45	397.01	high	147.4±15.0	4.5×10^{-10}	1.17±0.43
2	0.44	109.03	low	-56.4±15.3	4.7×10^{-10}	-1.01±0.42	14	0.49	548.94	low	142.9±16.1	8.2×10^{-6}	1.07±0.42
3	0.44	125.03	low	-78.0±23.7	2.2×10^{-10}	-1.06±0.42	15	0.49	604.03	very low	-81.3±31.3	5.3×10^{-9}	-0.87±0.41
4	0.44	291.01	low	-78.8±34.4	1.7×10^{-9}	-0.74±0.41	16	0.49	526.96	low	83.0±12.2	2.3×10^{-8}	0.98±0.42
5	0.44	207.06	low	-90.9±38.4	1.1×10^{-10}	-0.85±0.41	17	0.49	423.01	low	129.7±17.5	1.5×10^{-8}	0.93±0.42
6	0.44	163.06	low	-79.7±30.7	3.7×10^{-10}	-0.84±0.41	18	0.49	681.99	low	143.6±22.1	1.7×10^{-9}	0.78±0.41
7	0.44	441.99	low	169.6±19.6	1.0×10^{-8}	0.96±0.42	19	0.49	313.95	low	103.3±14.3	1.2×10^{-7}	0.98±0.42
8	0.45	508.58	low	171.8±18.1	1.8×10^{-11}	1.05±0.42	20	0.49	343.04	high	64.9±12.2	8.4×10^{-5}	0.82±0.41
9	0.45	506.58	low	178.1±18.9	3.5×10^{-6}	1.03±0.42	21	0.49	446.99	high	75.9±13.8	2.7×10^{-7}	0.82±0.41
10	0.45	413	low	121.8±13.8	4.8×10^{-7}	1.13±0.43	22	0.52	383.05	high	-73.2±35.0	1.4×10^{-7}	-0.64±0.41
11	0.45	398.01	low	171.3±16.0	7.7×10^{-11}	1.19±0.43	23	0.52	284.96	low	-82.5±46.2	3.2×10^{-6}	-0.59±0.40
12	0.45	397.01	high	147.4±15.0	5.3×10^{-8}	1.17±0.43	24	0.52	465	low	136.7±17.5	8.6×10^{-7}	0.96±0.42

Table A14 (continued)

Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI	Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI
25	0.53	206.99	high	-70.3±23.4	1.2×10^{-5}	-0.91±0.42	49	0.75	324.02	low	120.4±20.1	2.0×10^{-9}	0.78±0.41
26	0.55	192.98	high	-62.5±16.9	5.9×10^{-12}	-1.06±0.42	50	0.75	438.9	very low	-83.2±33.2	1.0×10^{-8}	-0.84±0.41
27	0.56	399.04	low	-67.6±25.4	1.5×10^{-6}	-0.79±0.41	51	0.75	267.75	very low	100.5±27.4	6.8×10^{-8}	0.50±0.40
28	0.56	692.14	very low	93.3±13.9	3.2×10^{-6}	0.94±0.42	52	0.76	560.99	very low	-70.8±24.8	2.1×10^{-8}	-0.87±0.41
29	0.56	226.81	low	164.2±14.7	6.6×10^{-3}	1.26±0.43	53	0.76	545	low	-56.5±28.0	7.0×10^{-8}	-0.56±0.40
30	0.58	224.81	low	229.8±21.8	1.1×10^{-7}	1.01±0.42	54	0.76	426.98	low	154.7±20.6	5.3×10^{-6}	0.88±0.4
31	0.6	124.99	low	-74.4±55.6	1.4×10^{-7}	-0.42±0.40	55	0.77	218.98	high	-30.8±10.4	1.1×10^{-4}	-0.70±0.41
32	0.64	261.01	very high	-69.0±20.7	2.1×10^{-8}	-1.00±0.42	56	0.77	275.02	high	-59.0±34.7	1.4×10^{-9}	-0.48±0.40
33	0.65	245.01	high	-58.1±29.5	4.1×10^{-13}	-0.55±0.40	57	0.77	573.03	low	-61.8±37.0	1.5×10^{-8}	-0.48±0.40
34	0.68	185.82	very low	79.0±12.9	5.3×10^{-6}	0.90±0.42	58	0.78	603.04	very low	-62.5±38.6	2.5×10^{-9}	-0.47±0.40
35	0.68	274.18	very low	-71.8±22.5	1.0×10^{-8}	-0.98±0.42	59	0.78	633.05	low	-90.7±47.4	1.3×10^{-11}	-0.68±0.41
36	0.69	419.68	very low	-81.1±59.2	6.4×10^{-3}	-0.45±0.40	60	0.78	540.02	low	-74.6±30.4	6.2×10^{-8}	-0.77±0.41
37	0.69	309.71	low	95.7±17.6	6.2×10^{-10}	0.76±0.41	61	0.78	406.97	very low	-92.8±41.4	3.5×10^{-7}	-0.81±0.41
38	0.69	416.6	very low	719.7±40.7	1.4×10^{-7}	0.80±0.41	62	0.84	305.03	high	-88.6±31.0	2.0×10^{-9}	-1.00±0.42
39	0.69	268.76	low	387.5±31.1	4.6×10^{-6}	0.89±0.42	63	0.84	589.02	very low	-82.3±37.8	1.2×10^{-8}	-0.72±0.41
40	0.69	414.6	very low	341.1±28.5	2.0×10^{-6}	0.92±0.42	64	0.85	575.01	low	-79.8±33.5	1.4×10^{-10}	-0.78±0.41
41	0.69	311.7	low	80.1±16.9	2.0×10^{-4}	0.70±0.41	65	0.85	556.99	very low	-82.6±27.4	1.3×10^{-8}	-1.01±0.42
42	0.7	169.85	very low	60.9±14.0	3.1×10^{-4}	0.68±0.41	66	0.9	516.97	low	-78.0±27.0	1.1×10^{-6}	-0.93±0.42
43	0.71	203.82	low	143.3±21.5	3.9×10^{-6}	0.80±0.41	67	0.93	246.99	very high	-63.0±15.8	3.0×10^{-11}	-1.15±0.43
44	0.71	204.98	high	-28.1±35.1	1.6×10^{-5}	-0.19±0.40	68	0.98	188.99	very high	-33.3±13.7	4.9×10^{-3}	-0.58±0.40
45	0.72	270.76	low	540.9±43.1	1.5×10^{-10}	0.71±0.41	69	1.19	245.07	high	44.8±8.4	2.3×10^{-8}	0.89±0.42
46	0.72	343.67	low	224.6±38.7	1.6×10^{-10}	0.57±0.40	70	1.33	238.06	low	121.5±12.4	1.5×10^{-7}	1.26±0.43
47	0.74	571.02	very low	-74.0±23.8	1.7×10^{-7}	-0.97±0.42	71	1.4	303.02	high	-71.5±19.7	6.0×10^{-8}	-1.12±0.43
48	0.74	232.98	high	-78.2±25.4	3.3×10^{-4}	-0.99±0.42	72	1.49	284.09	high	61.4±9.1	5.3×10^{-8}	1.05±0.42

Table A14 (continued)

Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI	Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI
73	1.52	317.06	high	60.9±8.0	6.2×10^{-8}	1.20±0.43	99	3.24	258.82	low	142.1±18.6	5.6×10^{-10}	0.92±0.42
74	1.53	322.07	low	96.0±9.8	8.0×10^{-7}	1.36±0.44	100	3.28	194.05	very high	-47.2±20.7	4.7×10^{-5}	-0.60±0.40
75	1.56	186.04	very low	130.0±12.5	9.7×10^{-5}	1.30±0.44	101	3.29	224.06	high	-69.2±16.8	6.7×10^{-6}	-1.24±0.43
76	1.57	324.03	low	82.1±11.8	5.5×10^{-7}	1.01±0.42	102	3.29	256.82	low	178.4±19.1	5.0×10^{-7}	1.02±0.42
77	1.79	317.01	very low	-73.5±28.1	3.2×10^{-8}	-0.81±0.41	103	3.3	267.85	low	95.7±12.9	4.4×10^{-8}	1.03±0.42
78	2.07	289.16	low	-71.0±24.6	4.7×10^{-10}	-0.88±0.41	104	3.38	189.2	very low	43.4±6.2	1.4×10^{-4}	1.17±0.43
79	2.76	209.84	low	261.3±31.6	3.4×10^{-8}	0.75±0.41	105	3.38	432	low	123.6±15.5	7.4×10^{-14}	1.01±0.42
80	2.82	178.05	very high	-33.5±11.0	2.0×10^{-3}	-0.73±0.41	106	3.51	295.13	high	-48.3±10.3	6.2×10^{-12}	-1.23±0.43
81	2.82	360.13	low	-77.3±22.7	9.6×10^{-7}	-1.09±0.42	107	3.52	425.16	low	-67.8±17.7	8.1×10^{-13}	-1.14±0.43
82	2.83	441.06	low	-64.6±14.0	6.0×10^{-5}	-1.35±0.44	108	3.53	439.16	low	-77.3±18.2	3.6×10^{-4}	-1.36±0.44
83	2.83	134.28	low	-75.5±22.9	1.1×10^{-5}	-1.04±0.42	109	3.55	447.09	low	-64.6±41.9	1.0×10^{-4}	-0.45±0.40
84	2.84	381.11	very low	-68.5±20.9	3.3×10^{-6}	-0.98±0.42	110	3.63	276.05	low	-59.7±14.3	3.3×10^{-5}	-1.17±0.43
85	2.85	400.15	low	-70.8±59.5	2.3×10^{-7}	-0.36±0.40	111	3.65	387.17	low	-50.1±19.6	1.7×10^{-4}	-0.68±0.41
86	2.85	367.23	low	-88.5±36.0	1.6×10^{-9}	-0.86±0.41	112	3.66	298.14	very low	-73.2±21.4	1.2×10^{-5}	-1.06±0.42
87	2.86	399.14	low	-67.6±53.5	4.0×10^{-8}	-0.37±0.40	113	3.72	383.16	low	-65.0±17.2	2.8×10^{-9}	-1.10±0.42
88	2.86	457.05	low	-78.8±24.0	1.6×10^{-7}	-1.06±0.42	114	3.74	657.2	very low	-41.7±28.3	2.0×10^{-9}	-0.38±0.40
89	2.94	184.1	high	-56.3±18.4	7.1×10^{-9}	-0.84±0.41	115	3.75	263.1	very high	34.7±11.7	9.8×10^{-3}	0.52±0.40
90	3.01	277.86	low	189.0±19.3	1.6×10^{-5}	1.05±0.42	116	3.76	330.05	low	-74.0±34.2	3.9×10^{-12}	-0.67±0.41
91	3.02	241.89	low	178.6±19.9	6.5×10^{-8}	0.99±0.42	117	3.76	367.12	low	-93.9±84.7	7.0×10^{-10}	-0.41±0.40
92	3.02	279.86	low	175.5±19.1	1.4×10^{-9}	1.02±0.42	118	3.77	835.31	low	65.0±13.5	6.4×10^{-4}	0.75±0.41
93	3.03	281.86	low	212.7±20.7	8.0×10^{-7}	1.04±0.42	119	3.77	417.22	low	-78.1±27.2	4.4×10^{-7}	-0.92±0.42
94	3.04	243.88	low	138.8±19.3	4.4×10^{-8}	0.88±0.41	120	3.79	385.19	low	-73.8±51.3	5.8×10^{-11}	-0.45±0.40
95	3.06	265.12	low	-83.5±22.9	8.0×10^{-7}	-1.22±0.43	121	3.81	431.21	low	-66.9±19.9	2.0×10^{-8}	-0.99±0.42
96	3.09	223.84	low	145.3±18.8	1.4×10^{-10}	0.93±0.42	122	3.83	405.13	high	44.3±10.1	6.1×10^{-4}	0.73±0.41
97	3.14	250.07	low	-55.5±24.5	1.8×10^{-9}	-0.63±0.41	123	3.85	389.22	low	-71.7±24.7	1.5×10^{-5}	-0.89±0.42
98	3.22	253.83	very low	161.4±19.4	2.8×10^{-6}	0.95±0.42	124	3.85	401.18	low	-70.3±24.0	1.9×10^{-8}	-0.89±0.42

Table A14 (continued)

Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI	Feature No.	RT (min)	*m/z*	feature intensity	% variation	*p*-value	effect size ±CI
125	3.89	397.1	low	-72.6±34.2	4.6×10^{-9}	-0.66±0.41	137	4.6	378.11	low	79.2±12.5	9.3×10^{-11}	0.93±0.42
126	3.89	399.17	low	-69.7±27.0	2.2×10^{-7}	-0.78±0.41	138	4.61	449.21	low	-56.4±15.4	8.5×10^{-11}	-1.01±0.42
127	3.9	205.13	very low	-76.9±29.8	1.1×10^{-9}	-0.82±0.41	139	4.61	401.15	low	-63.4±17.9	2.1×10^{-12}	-1.02±0.42
128	3.9	483.14	very low	-59.3±24.4	5.5×10^{-9}	-0.69±0.41	140	4.61	394.13	high	261.2±22.7	1.0×10^{-4}	1.04±0.42
129	3.91	467.07	low	-84.0±41.1	1.5×10^{-9}	-0.69±0.41	141	4.61	406.19	low	-73.7±22.9	3.9×10^{-9}	-1.00±0.42
130	4.03	499.21	very low	-80.6±25.8	3.9×10^{-10}	-1.02±0.42	142	4.61	365.14	low	63.5±6.1	3.4×10^{-11}	1.62±0.46
131	4.33	371.06	low	-65.2±18.4	4.6×10^{-9}	-1.04±0.42	143	5.01	282.12	low	120.4±13.2	1.2×10^{-6}	1.16±0.43
132	4.35	471.19	low	-66.1±23.7	5.6×10^{-10}	-0.82±0.41	144	5.46	284.12	low	-48.3±13.9	3.9×10^{-8}	-0.91±0.42
133	4.41	299.1	low	-55.1±13.7	3.2×10^{-9}	-1.10±0.42	145	5.47	302.13	low	-51.3±12.7	2.5×10^{-8}	-1.08±0.42
134	4.43	282.01	high	31.8±4.4	9.2×10^{-12}	1.26±0.43	146	5.79	345.15	low	-85.0±38.4	3.9×10^{-9}	-0.75±0.41
135	4.54	253.12	low	-62.5±14.5	4.1×10^{-9}	-1.24±0.43	147	6.43	201.13	high	45.7±6.1	5.5×10^{-6}	1.24±0.43
136	4.59	261.03	low	54.9±8.5	4.1×10^{-12}	1.05±0.42							